Bayesian Indirect Inference using a Parametric Auxiliary Model

C. C. Drovandi and A. N. Pettitt
School of Mathematical Sciences
Queensland University of Technology, Brisbane, Australia 4001

A. Lee
Department of Statistics
Warwick University, Coventry, UK CV4 7AL

email: c.drovandi@qut.edu.au

April 29, 2014

Abstract

Indirect inference (II) is a methodology for estimating the parameters of an intractable (generative) model on the basis of an alternative parametric (auxiliary) model that is both analytically and computationally easier to deal with. Such an approach has been well explored in the classical literature but has received substantially less attention in the Bayesian paradigm. The purpose of this paper is to compare and contrast a collection of what we call parametric Bayesian indirect inference (pBII) methods. One class of pBII methods uses approximate Bayesian computation (referred to here as ABC II) where the summary statistic is formed on the basis of the auxiliary model, using ideas from II. Another approach proposed in the literature, referred to here as parametric Bayesian indirect likelihood (pBIL), we show to be a fundamentally different approach to ABC II. We devise new theoretical results for pBIL to give extra insights into its behaviour and also its differences with ABC II. Furthermore, we examine in more detail the assumptions required to use each pBII method. The results, insights and comparisons developed in this paper are illustrated on simple examples and two other substantive applications. The first of the substantive examples involves performing inference for complex quantile distributions based on simulated data while the second is for estimating the parameters of a trivariate stochastic process describing the evolution of macroparasites within a host based on real data. We create a novel framework called Bayesian indirect likelihood (BIL) which encompasses pBII as well as general ABC methods so that the connections between the methods can be established.

KEYWORDS: approximate Bayesian computation, likelihood-free methods, Markov jump processes, quantile distributions, simulated likelihood

1 Introduction

Approximate Bayesian computation (ABC) now plays an important role in performing (approximate) Bayesian inference for the parameter of a proposed statistical model (called the generative model here) that has an intractable likelihood. Despite the intense attention ABC has recently received, the approach still suffers from several drawbacks. An obvious disadvantage is the usual necessity to reduce the data to a low dimensional summary statistic. This leads to a loss of information that is difficult to quantify. The second, often less severe but sometimes related, drawback is the computational challenge of achieving stringent matching between the observed and simulated summary statistics.

In situations where an alternative parametric model (referred to as an auxiliary model) can be formulated that has a tractable likelihood, the methodology known as indirect inference (II) (see, for example, Gourieroux et al. (1993) and Heggland and Frigessi (2004)) is applicable. II has been thoroughly examined in the classical framework. Most methods differ in the way that observed and simulated data are compared via the auxiliary model. We expand on this later in the article. For the moment we note that some key references are Gourieroux et al. (1993); Smith (1993); Gallant and Tauchen (1996).

However, II has been far less studied in the Bayesian paradigm. Drovandi et al. (2011) developed an ABC approach that uses II to obtain summary statistics. In particular, the estimated parameter of the auxiliary model fitted to the data becomes the observed summary statistic. We adopt a similar naming convention to Gleim and Pigorsch (2013) and refer to this method as ABC IP where 'I' stands for 'indirect' and 'P' stands for 'parameter'. Gleim and Pigorsch (2013) also list another method, ABC IL (where 'L' stands for 'likelihood'), which is essentially an ABC version of Smith (1993). This approach follows ABC IP in the sense that the parameter estimate of the auxiliary model is again the summary statistic. However, the ABC discrepancy is based on the auxiliary likelihood, rather than a direct comparison of the auxiliary parameters.

Gleim and Pigorsch (2013) advocate a slightly different approach to ABC with II, which is effectively an ABC version of the classical approach in Gallant and Tauchen (1996). Here Gleim and Pigorsch (2013) use the score vector based on the auxiliary model as the summary statistic, which is referred to as ABC IS (here 'S' stands for 'score'). The parameter value used in the score is given by the MLE of the auxiliary model fitted to the observed data. This approach can be far cheaper from a computational point of view since it avoids an expensive fitting of the auxiliary model to each dataset simulated from the generative model required in ABC IP and ABC IL.

Throughout the paper the collection of approaches that use the parametric auxiliary model to form summary statistics is referred to as ABC II methods. An advantage of this approach over more traditional summary statistics is one can check the usefulness of the II summary statistic prior to the ABC analysis. Additionally, if the auxiliary model is parsimonious, then the summary statistic can be low-dimensional.

Gallant and McCulloch (2009) (see also Reeves and Pettitt (2005)) suggest an alternative approach for combining II with Bayesian inference. This method has similar steps to ABC IP and ABC IL but essentially uses the likelihood of the auxiliary model as a replacement to the intractable generative model likelihood. We note here that this is a fundamentally

different approach as it is not a standard ABC method. In particular, there is no comparison of summary statistics and no need to choose an ABC tolerance. Here we refer to this method as parametric Bayesian indirect likelihood (pBIL). The focus of this paper is the application of a parametric auxiliary model for the full data, which we refer to as pdBIL (where 'd' stands for 'data'). However, the ideas in this paper are equally applicable if a parametric model is applied at the summary statistic (not necessarily obtained using ABC II techniques) level (i.e. some data reduction technique has been applied, see Blum et al. (2013) for a review). This is referred to as psBIL (where 's' stands for 'summary statistic'). We show that the Bayesian version of the synthetic likelihood method of Wood (2010) is a psBIL method. In the paper we refer to the collection of ABC II and pBIL approaches as pBII methods ('B'ayesian 'I'ndirect 'I'nference using a 'p'arametric auxiliary model).

In the process of reviewing these pBII methods, we create a novel framework called Bayesian indirect likelihood (BIL) which encompasses pBII as well as ABC methods generally. In particular, if a specific non-parametric auxiliary model is selected (npBIL) instead of a parametric one (pBIL), then the general ABC method is recovered. The non-parametric kernel can be applied either at the full data (npdBIL) or summary statistic (npsBIL) level. The ABC II approaches are thus a special type of npsBIL method where the summary statistic is formed on the basis of a parametric auxiliary model. This framework is shown in Figure 1, which also highlights the methods that this paper addresses.

This article does not develop any new algorithms for pBII. However, this paper does make several interesting and useful contributions. Firstly, we explore the pdBIL method in more detail theoretically, and recognise that it is fundamentally different to ABC II. The behaviour of this method is also substantially different. A technique sometimes applied with classical II methods is to increase the simulated dataset size beyond that of the observed data in order to reduce the variability of the estimated quantities under the auxiliary model (see, for example, Smith (1993); Gourieroux et al. (1993); Gallant and McCulloch (2009)). We demonstrate that pdBIL and ABC II behave differently for increasing size of the simulated data. Our second contribution is to compare the assumptions required for each pBII approach. Our theoretical and empirical results indicate that the pBIL method will provide good approximations if the auxiliary model is sufficiently flexible to provide a good fit to data simulated from the true model based on parameter values within regions of non-negligible posterior probability. ABC II methods rely on the parameter estimate or the score of the auxiliary model to provide a near-sufficient summary statistic. Finally, our creation of the general BIL framework provides a clear way to see the connections between pBII and other methods.

The paper is organised as follows. In Section 2 the notation used throughout the paper is defined. The ABC II methods are reviewed in Section 3. The pBIL approach is presented in Section 4. The theoretical developments in this section, which offer additional insight into the pBIL approximation, are new. In addition, this section demonstrates how the synthetic likelihood approach of Wood (2010) is a pBIL method on the summary statistic level. Section 5 shows how ABC can be recovered as a BIL method via a non-parametric choice of the auxiliary model. Section 6 provides a comparison between ABC II and pdBIL. The contributions of this article are demonstrated on examples with varying complexity in Section 7. The highlights of this section include improved approximate inferences for quantile distributions and a multivariate Markov jump process explaining the evolution of parasites within a host. The article concludes with a discussion in Section 8.

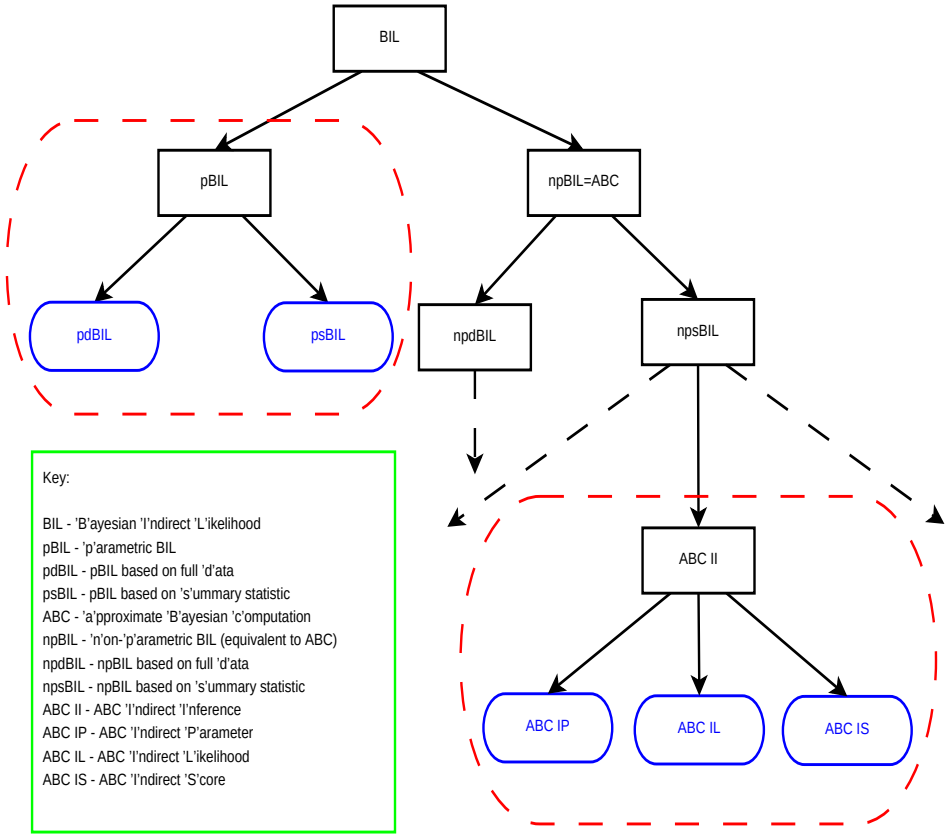


Figure 1: The general BIL Framework. Here the rounded rectangles indicate particular instances of methods. The dashed arrows indicate that there are multiple instances of that class of method in the literature that are not placed on this diagram. The dashed larger rounded rectangles indicate the methods that this paper focusses on. That is, BII methods that make use of, in some way, a parametric auxiliary model (so-called pBII methods).

2 Notation

Consider an observed dataset $\boldsymbol{y}$ taking values in Y of dimension N assumed to have arisen from a generative model with an intractable likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is the parameter of this model. Intractability here refers to the inability to compute $p(\boldsymbol{y}|\boldsymbol{\theta})$ pointwise as a function of $\boldsymbol{\theta}$. We assume that there is a second statistical model that has a tractable likelihood function. We denote the likelihood of this auxiliary model by $p_A(\boldsymbol{y}|\boldsymbol{\phi})$, where $\boldsymbol{\phi} \in \Phi$ denotes the parameter of this auxiliary model. There does not necessarily need to be any obvious connection between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The auxiliary model could be purely a data analytic model that does not offer any mechanistic explanation of how the observed data arose. The parameter estimate of the auxiliary model when fitted to the observed data is given by $\boldsymbol{\phi}(\boldsymbol{y})$. Assuming a prior distribution on the parameters of the generative model, $p(\boldsymbol{\theta})$, our interest is in sampling from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, or some approximation thereof.

We denote data simulated from the model as $\boldsymbol{x} \in \mathsf{Y}$. In this paper, we also consider the effect of using n independent replicates of data simulated from the model, which we denote $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and define $p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\theta})$. Therefore the total size of $\boldsymbol{x}_{1:n}$ is nN. Note that this could also relate to a stationary time series simulation of length nN (see for example Gallant and McCulloch (2009)). In the case of a stationary time series or independent and identically distributed (iid) data it is not a requirement for the simulated dataset size to be a multiple of the observed data size. However for the sake of simplicity we restrict n to be a positive integer.

Since the likelihood of the auxiliary model is tractable, we can potentially consider richly parameterised statistical models to capture the essential features of the data. We assume throughout the paper that the dimensionality of the auxiliary model parameter is at least as large as the dimensionality of the generative model parameter, i.e. $\dim(\boldsymbol{\phi}) \geq \dim(\boldsymbol{\theta})$. Jiang and Turnbull (2004) note that it still may be possible to obtain useful estimates of a subset of the parameters when this assumption does not hold. We do not consider this here.

3 Approximate Bayesian Computation with Indirect Inference

3.1 Approximate Bayesian Computation

ABC is widely becoming a standard tool for performing (approximate) Bayesian inference on statistical models with computationally intractable likelihood evaluations but where simulation is straightforward. ABC analyses set $n = 1$ so that the simulated dataset is the same size as the observed. However, for the purposes of this paper, we relax this commonly applied restriction for the moment.

In ABC, a summary statistic is defined by a collection of functions $s_n : \mathsf{Y}^n \to \mathsf{S}$, for each $n \in \mathbb{N}$. Henceforth, the subscript n is omitted to ease presentation. Proposed parameter values that produce a simulated summary statistic, $s(\boldsymbol{x}_{1:n})$, 'close' to the observed summary statistic, $s(\boldsymbol{y})$, are given more weight. Here we define 'close' by the discrepancy function $\rho(s(\boldsymbol{x}_{1:n}), s(\boldsymbol{y}))$ and a kernel weighting function, $K_\epsilon(\rho(s(\boldsymbol{x}_{1:n}), s(\boldsymbol{y})))$, where ϵ is a bandwidth

referred to as the ABC tolerance. The ABC target distribution is given by

$$p_{\epsilon,n}(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p_{\epsilon,n}(\boldsymbol{y}|\boldsymbol{\theta}), \tag{1}$$

where

$$p_{\epsilon,n}(\boldsymbol{y}|\boldsymbol{\theta}) \propto \int_{\mathsf{Y}^n} p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta})K_\epsilon(\rho(s(\boldsymbol{x}_{1:n}), s(\boldsymbol{y})))d\boldsymbol{x}_{1:n},$$

is referred to here as the ABC likelihood. It can be shown that if $s(\cdot)$ is a sufficient statistic and $n = 1$ then $p_{\epsilon,1}(\boldsymbol{\theta}|\boldsymbol{y}) \to p(\boldsymbol{\theta}|\boldsymbol{y})$ as $\epsilon \to 0$ (Blum, 2010). Unfortunately, ABC cannot be trusted as the value of n is increased in the sense that the target $p_{\epsilon,n}(\boldsymbol{\theta}|\boldsymbol{y})$ can move further away from $p(\boldsymbol{\theta}|\boldsymbol{y})$. In the simple example in Appendix A, the posterior distribution for a univariate θ converges to a point mass centred on the single observation y as $n \to \infty$ and $\epsilon \to 0$ (see also Drovandi (2012, pp. 28-29) for a similar example). We also verify this behaviour empirically on a toy example. This suggests that whilst it is tempting to increase the simulated dataset size to reduce the variability of the simulated summary statistic, such an approach is fraught with danger.

Standard ABC procedures can make use of n simulated datasets but in a different way. Here the ABC likelihood is estimated via $1/n \sum_{i=1}^{n} K_\epsilon(\rho(s(\boldsymbol{x}_i), s(\boldsymbol{y})))$. However, we note that this is an unbiased estimate (regardless of n) of a more standard ABC likelihood and thus this process does not alter the ABC approximation (Andrieu and Roberts, 2009). Therefore, for the remaining presentation in this section we set $n = 1$.

How the parameter value is proposed depends on the chosen ABC algorithm. Here we use Markov chain Monte Carlo (MCMC ABC, Marjoram et al. (2003)) with the proposal distribution q chosen carefully, and sometimes based on previous ABC analyses. The approach is shown in Algorithm 1 for completeness. In the algorithm, T is the number of iterations. We choose $\boldsymbol{\theta}^0$, the initial value of the chain, so that it is well supported by the target distribution so we do not use any burn-in. This algorithm is mainly used for simplicity, although it is important to note that other ABC algorithms could be applied.

Algorithm 1 MCMC ABC algorithm of Marjoram et al. (2003).

Set $\boldsymbol{\theta}^0$
for $i = 1$ **to** T **do**
  Draw $\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{i-1})$
  Simulate $\boldsymbol{x}^* \sim p(\cdot|\boldsymbol{\theta}^*)$
  Compute $r = \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)K_\epsilon(\rho(s(\boldsymbol{x}^*),s(\boldsymbol{y})))}{p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})K_\epsilon(\rho(s(\boldsymbol{x}^{i-1}),s(\boldsymbol{y})))}$
  if uniform$(0, 1) < r$ **then**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$ and $\boldsymbol{x}^i = \boldsymbol{x}^*$
  else
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$ and $\boldsymbol{x}^i = \boldsymbol{x}^{i-1}$
  end if
end for

One difficult aspect in ABC is that often some form of data reduction is required, to avoid the curse of dimensionality associated with comparing all the data at once. This choice of

summary statistic is therefore crucial for a good approximation, even when ϵ can be reduced to practically 0. In some applications it is possible to propose another model that provides a good description of the data. The summary statistic used in ABC can be formulated based on this auxiliary model. These approaches are summarised below.

3.2 ABC IP

Drovandi et al. (2011) suggest using the parameter estimate of the auxiliary model as the summary statistic to use in ABC. Data are simulated from the generative model based on a proposed parameter value, then the auxiliary model parameter is estimated based on this simulated data. The way the auxiliary parameter is estimated provides a mapping between the generative model and the auxiliary model parameters. The ABC algorithm uses a noisy mapping between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ through the simulated data, $\boldsymbol{x}$, generated on the basis of $\boldsymbol{\theta}$, $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x})$. For this purpose (explained below) we use the maximum likelihood estimate (MLE) of the auxiliary model

$$\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x}) = \arg\max_{\boldsymbol{\phi} \in \Phi} p_A(\boldsymbol{x}|\boldsymbol{\phi}).$$

where $\boldsymbol{x} \sim p(\cdot|\boldsymbol{\theta})$. ABC IP relies on the following assumption:

Assumption 1 (*ABC IP Assumptions*)**.** *The estimator of the auxiliary parameter, $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x})$, is unique for all $\boldsymbol{\theta}$ with positive prior support.*

It is important to note that ABC IP (as well as other ABC approaches below) use $n = 1$ so the approximation quality of the method can depend on the statistical efficiency of the estimator $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x})$ based on this finite sample. Additionally, the MLE is asymptotically sufficient (Cox and Hinkley, 1979, pp. 307). For this reason we advocate the use of the MLE in general as it is typically more efficient than other estimators like sample moments (for the auxiliary model). Sample moments may be computationally easier to obtain, but are likely to result in a poorer ABC approximation if the statistical efficiency is lower than the MLE. We note that the optimal choice of auxiliary estimator (trading off between computational effort and statistical efficiency) may be problem dependent. An additional complication is that the auxiliary model is fitted based on data generated from a different model, $\boldsymbol{x} \sim p(\cdot|\boldsymbol{\theta})$. Therefore, the efficiency of $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\theta})$ should be based on $p(\boldsymbol{x}|\boldsymbol{\theta})$ not $p_A(\boldsymbol{x}|\boldsymbol{\phi})$ (see, for example, Cox (1961)). This can be investigated by simulation.

In Section 7.3, we provide an example where the auxiliary model does not satisfy Assumption 1, creating difficulties for ABC IP. The ABC II methods (and the pdBIL method) that follow do not necessarily require unique auxiliary parameter estimates.

An advantage of the II approach to obtaining summary statistics is that the summary statistic will likely be useful if the auxiliary model fits the observed data (see Section 3.5 for more discussion). In the case of ABC IP, the (approximate) covariance matrix of the auxiliary parameter estimate based on the observed data can be estimated by the inverse of the observed information matrix (we denote this information matrix by $\boldsymbol{J}(\boldsymbol{\phi}(\boldsymbol{y}))$). Intuitively, we expect this discrepancy function to be more efficient than a Euclidean distance, as it can take into account the variability of summary statistics and the correlations between summary statistics.

Denoting the observed summary statistic as $\boldsymbol{\phi}(\boldsymbol{y})$ and the simulated summary statistic as $\boldsymbol{\phi}(\boldsymbol{x})$ (dropping $\boldsymbol{\theta}$ for notational convenience), we use the following discrepancy for ABC IP

$$\rho(s(\boldsymbol{x}), s(\boldsymbol{y})) = \sqrt{(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{y}))^T \boldsymbol{J}(\boldsymbol{\phi}(\boldsymbol{y}))(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{y}))}.$$

It is important to note that this is essentially an ABC version of the classical approach in Gourieroux et al. (1993). A more appropriate weighting matrix may involve considering the variance of $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\theta})$ when the data are generated under an alternative model, $\boldsymbol{x} \sim p(\cdot|\boldsymbol{\theta})$ (Cox (1961) provides a result), the so-called sandwich estimator.

3.3 ABC IL

Gleim and Pigorsch (2013) also describe an approach that uses the auxiliary likelihood to set up an ABC discrepancy. Here the ABC discrepancy is

$$\rho(s(\boldsymbol{x}), s(\boldsymbol{y})) = \log p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{y})) - \log p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{x})).$$

This is effectively an ABC version of the classical approach of Smith (1993). We note that $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{y}))$ will remain unchanged throughout the algorithm and provides an upperbound for values of $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{x}))$ obtained for every simulated dataset. ABC IL uses the same summary statistic as ABC IP but uses a discrepancy based on the likelihood rather than the Mahalanobis distance. Note that the discrepancy function for ABC IP appears in the second order Taylor series approximation of $\log p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{x}))$ about $\log p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{y}))$ (Davison, 2003, pp. 126) assuming standard regularity conditions for $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{y}))$ and $\boldsymbol{\phi}(\boldsymbol{y})$. The ABC tolerance could be viewed here as a certain cut-off value of the auxiliary log-likelihood.

The ABC IL approach relies on the following assumption:

Assumption 2 (*ABC IL Assumptions*)**.** *The auxiliary likelihood evaluated at the auxiliary estimate, $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\theta}))$, is unique for all $\boldsymbol{\theta}$ with positive prior support.*

We note that this assumption can still be satisfied even when the auxiliary model does not have a unique MLE (see Section 7.3 for an example).

The ABC IP and ABC IL methods use parameter estimates of the auxiliary model as summary statistics and can thus be expensive as it can involve a numerical optimisation every time data is simulated from the generative model. The next approach to obtaining summary statistics from II avoids this optimisation step.

3.4 ABC IS

Gleim and Pigorsch (2013) advocate the use of the score vector of the auxiliary model evaluated at the auxiliary MLE, $\boldsymbol{\phi}(\boldsymbol{y})$, as the summary statistic. We denote the score vector of the auxiliary model as

$$\boldsymbol{S}_A(\boldsymbol{y}, \boldsymbol{\phi}) = \left(\frac{\partial \log p_A(\boldsymbol{y}|\boldsymbol{\phi})}{\partial \phi_1}, \cdots, \frac{\partial \log p_A(\boldsymbol{y}|\boldsymbol{\phi})}{\partial \phi_{dim(\boldsymbol{\phi})}} \right)^T,$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{dim(\boldsymbol{\phi})})^T$.

Each component of the summary statistic involving the observed data and the MLE, $\boldsymbol{S}_A(\boldsymbol{y}, \boldsymbol{\phi}(\boldsymbol{y}))$, is numerically 0. Thus the search is for parameter values of the generative model that lead to simulated data, $\boldsymbol{x}$, that produces a score close to $\mathbf{0}$. Noting that the approximate covariance matrix of the observed score is given by $\boldsymbol{J}(\boldsymbol{\phi}(\boldsymbol{y}))$, the following ABC discrepancy is obtained for ABC IS

$$\rho(s(\boldsymbol{x}), s(\boldsymbol{y})) = \sqrt{\boldsymbol{S}_A(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{y}))^T \boldsymbol{J}(\boldsymbol{\phi}(\boldsymbol{y}))^{-1} \boldsymbol{S}_A(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{y}))}.$$

This is essentially an ABC version of Gallant and Tauchen (1996).

This approach is fast relative to ABC IP when the MLE of the auxiliary model is not analytic whilst the score is analytic since no numerical optimisation is required every time data are simulated from the generative model. Of course, it may be necessary to estimate the score numerically, which would add another layer of approximation and may be slower. In the examples of this paper we are able to obtain the score analytically. ABC IS relies on the following assumptions:

Assumption 3 (*ABC IS Assumptions*)**.** *The MLE of the auxiliary model fitted to the observed data, $\boldsymbol{\phi}(\boldsymbol{y})$, is an interior point of the parameter space of $\boldsymbol{\phi}$ and $\boldsymbol{J}(\boldsymbol{\phi}(\boldsymbol{y}))$ is positive definite. The log-likelihood of the auxiliary model, $\log p_A(\cdot|\boldsymbol{\phi})$, is differentiable and the score, $\boldsymbol{S}_A(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{y}))$, is unique for any $\boldsymbol{x}$ that may be drawn according to any $\boldsymbol{\theta}$ that has positive prior support.*

We note that this is a generally weaker assumption than ABC IP, since it may still hold even if the MLE of the auxiliary model is not unique.

3.5 Discussion on ABC II Summary Statistics

Only models in the exponential family possess a minimal sufficient statistic with dimension equal to that of $dim(\boldsymbol{\theta})$. For other models, under suitable conditions, the Pitman–Koopman–Darmois theorem states that the dimension of any sufficient statistic increases with the sample size. For many complex models, such as those considered in the ABC setting, the minimal sufficient statistic will be the full dataset (or the full set of order statistics if the data are iid). The summary statistic produced by ABC II will always have dimension $dim(\boldsymbol{\phi})$, and thus will not produce sufficient statistics in general (this argument of course carries over to any ABC method that uses some data reduction technique, see Blum et al. (2013) for a review). Intuitively, our suggestion is that the summary statistic produced by ABC II should carry most of the information contained in the observed data provided that the auxiliary model provides a good description of the data. Unfortunately, this is difficult to prove since it is usually not possible to quantify the amount of information lost in data reduction. Despite this, by conducting goodness-of-fit tests and/or residual analysis on the auxiliary model fit to the data will at least provide some guidance on the usefulness of the summary statistic produced by ABC II. This is in contrast to more traditional approach of summarising based on simple functions of the data (e.g. Drovandi and Pettitt (2011)), whose utility is difficult to assess prior to running an ABC analysis without performing an expensive simulation study. Furthermore, ABC II methods provide natural discrepancy functions between summary statistics

as shown above. Selecting the discrepancy function and determining appropriate weighting of the summary statistics in traditional ABC can be problematic.

It is well known that the choice of summary statistics in ABC involves a compromise between sufficiency and dimensionality (Blum et al., 2013). A low-dimensional and near-sufficient summary statistic represents an optimal trade-off. Another advantage of ABC II over usual ABC is the dimensionality of the ABC II summary statistic can be controlled by selecting parsimonious auxiliary models and using standard model choice techniques to choose between a set of possible auxiliary models (e.g. the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)).

4 Parametric Bayesian Indirect Likelihood (pBIL)

4.1 Parametric Bayesian Indirect Likelihood for the Full Data (pdBIL)

Reeves and Pettitt (2005) and Gallant and McCulloch (2009) propose a method that has similar steps to ABC IP and ABC IL but is theoretically quite different, as we show below. After data are simulated from the generative model, the auxiliary parameter is estimated. This auxiliary estimate is then passed into the auxiliary likelihood of the observed data. This likelihood is then treated in the usual way and fed into a Bayesian algorithm, for example MCMC. One first defines a collection of functions $\boldsymbol{\phi}_n : \Theta \times \mathsf{Y}^n \to \Phi$. The artificial likelihood is then defined as follows

$$p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathsf{Y}^n} p_A(\boldsymbol{y}|\boldsymbol{\phi}_n(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})) \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}_{1:n},$$

and the target distribution of this approach is given by

$$p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y}) \propto p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where the subscripts A and n denote the dependence of the target on the auxiliary model choice and the number of replicate simulated datasets, respectively. This approach is effectively a Bayesian version of the simulated quasi-maximum likelihood approach of Smith (1993). Smith (1993) proposes to maximise $p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Instead of applying this as an ABC discrepancy as in the ABC IL method above, Reeves and Pettitt (2005) and Gallant and McCulloch (2009) treat this auxiliary likelihood as a replacement to the likelihood of the generative model in the same way that Smith (1993) does.

It is important to note that this approach does not perform a comparison of summary statistics and hence there is no need to choose an ABC tolerance. Thus it is not a standard ABC algorithm. We refer to this approach simply as pdBIL, since we apply a 'p'arametric auxiliary model for the full 'd'ata. When the full data have been summarised as a summary statistic, $s(\boldsymbol{y})$, an alternative approach is to apply a parametric auxiliary model for the summary statistic likelihood, $p(s(\boldsymbol{y})|\boldsymbol{\theta})$ (see Section 4.2 for more details). In Figure 1, these methods fall under the pBIL class. Within this class, the focus of this paper is on the pdBIL method.

4.1.1 The pdBIL Approximation

The theoretical aspects of this approach have yet to be investigated in the literature. Some clues are offered in Reeves and Pettitt (2005) and Gallant and McCulloch (2009), but we formalise and extend the theory here. The subscript n is used to denote that this target remains an approximation to the true posterior distribution and that the approximate target may change with n (we show below that in general the target does depend on n). However, because it is not an ABC algorithm, it is unclear how pdBIL behaves as n increases. Gallant and McCulloch (2009) use a very large simulation size ($n \approx 700$), without a theoretical investigation.

We consider a potential limiting likelihood $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ with associated posterior

$$p_A(\boldsymbol{\theta}|\boldsymbol{y}) \propto p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))p(\boldsymbol{\theta}).$$

Note that the parameter of $p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta})$ is $\boldsymbol{\theta} \in \Theta$ but the parameter of $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ is $\boldsymbol{\phi}(\boldsymbol{\theta}) \in \Phi$.

To ease presentation, we define the random variable $\boldsymbol{\phi}_{\boldsymbol{\theta},n} = \boldsymbol{\phi}_n(\boldsymbol{\theta}, \boldsymbol{X}_{1:n})$ where $(\boldsymbol{X}_i)$ is a sequence of iid random variables distributed according to $p(\cdot|\boldsymbol{\theta})$, and we can write

$$p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) = \mathsf{E}\left[p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})\right],$$

where E is expectation with respect to the distribution of $\boldsymbol{X}_{1:n}$.

The results below provide sufficient conditions under which, as $n \to \infty$, $p_{A,n}(\cdot|\boldsymbol{y}) \to p_A(\cdot|\boldsymbol{y})$ pointwise and

$$\int_\Theta f(\boldsymbol{\theta})p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \to \int_\Theta f(\boldsymbol{\theta})p_A(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

where f is some function of $\boldsymbol{\theta}$ whose posterior expectation is of interest. Note that this does not assume that $p_A(\cdot|\boldsymbol{y}) = p(\cdot|\boldsymbol{y})$, which in general will not be the case.

A useful tool to allow us to answer both questions is provided by Billingsley (1999).

Theorem 1 (Billingsley (1999), Theorem 3.5)**.** *If X_n is a sequence of uniformly integrable random variables and X_n converges in distribution to X then X is integrable and $\mathsf{E}X_n \to \mathsf{E}X$.*

Remark 1. *A simple sufficient condition for uniform integrability is that for some $\delta > 0$,*

$$\sup_n \mathsf{E}\left(|X_n|^{1+\delta}\right) < \infty.$$

Result 1. *Assume that $p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) \to p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ as $n \to \infty$ for all $\boldsymbol{\theta}$ with positive prior support, $\inf_n \int_\Theta p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} > 0$ and $\sup_{\boldsymbol{\phi}\in\Phi} p_A(\boldsymbol{y}|\boldsymbol{\phi}) < \infty$. Then*

$$\lim_{n\to\infty} p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y}) = p_A(\boldsymbol{\theta}|\boldsymbol{y}).$$

Furthermore, if $f : \Theta \to \mathbb{R}$ is a continuous function satisfying $\sup_n \int_\Theta |f(\boldsymbol{\theta})|^{1+\delta} p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} < \infty$ for some $\delta > 0$ then

$$\lim_{n\to\infty} \int_\Theta f(\boldsymbol{\theta})p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} = \int_\Theta f(\boldsymbol{\theta})p_A(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.$$

Proof. The first part follows from the fact that the numerator of

$$p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_\Theta p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

converges pointwise and the denominator is positive and converges by the bounded convergence theorem. For the second part, if for each $n \in \mathbb{N}$, $\boldsymbol{\theta}_n$ is distributed according to $p_{A,n}(\cdot|\boldsymbol{y})$ and $\boldsymbol{\theta}$ is distributed according to $p_A(\cdot|\boldsymbol{y})$ then $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}$ in distribution as $n \to \infty$ by Scheffé's lemma (Scheffé, 1947). Since f is continuous, $f(\boldsymbol{\theta}_n)$ converges in distribution to $f(\boldsymbol{\theta})$ as $n \to \infty$ by the continuous mapping theorem and we conclude by application of Theorem 1. □

A simple condition for $p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) \to p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ as $n \to \infty$ to hold is provided by the following Result.

Result 2. *Assume that $p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})$ converges in probability to $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ as $n \to \infty$. If*

$$\sup_n \mathsf{E}\left[|p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n}|^{1+\delta}\right] < \infty$$

for some $\delta > 0$ then $p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) \to p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ as $n \to \infty$.

Proof. The result follows by application of Theorem 1. □

The function $\boldsymbol{\phi}(\boldsymbol{\theta})$ is often referred to as the mapping or binding function in the II literature. Assume that $\boldsymbol{\phi}_{\boldsymbol{\theta},n}$ converges in probability to $\boldsymbol{\phi}(\boldsymbol{\theta})$ and the function $\boldsymbol{\phi} : \Phi \to \mathbb{R}_+$ given by $\boldsymbol{\phi} \mapsto p_A(\boldsymbol{y}|\boldsymbol{\phi})$ is continuous. Then if $\sup_{\boldsymbol{\phi}\in\Phi} p_A(\boldsymbol{y}|\boldsymbol{\phi}) < \infty$, Result 2 holds. Therefore if the binding function is 1-1 and the auxiliary likelihood is bounded above then this is sufficient for Result 2. Gallant and McCulloch (2009) assume that the binding function is 1-1. However, the following demonstrates that a 1-1 mapping is not a necessary condition for Result 2. Assume that $p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})$ converges in probability to $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ and let $\sup_{\boldsymbol{\phi}\in\Phi} p_A(\boldsymbol{y}|\boldsymbol{\phi}) < \infty$. Then Result 2 holds. Thus, it suffices for $\boldsymbol{\theta}$ to produce a unique value for the auxiliary likelihood, $p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$, which can still be true even if the binding function is not 1-1. The following example where the auxiliary model is a mixture model demonstrates this principle.

Assume that the true model is $N(\boldsymbol{y};\theta,1)$ while the auxiliary model is a mixture of normals, $wN(\boldsymbol{y};\theta_1,1)+(1-w)N(\boldsymbol{y};\theta_2,1)$ so that $\boldsymbol{\phi} = (\theta_1,\theta_2,w)$. Assuming an infinite sample from the true model, there are an infinite number of MLEs of the auxiliary model; $\boldsymbol{\phi}(\theta) = (\theta,\theta,w)$ where $0 \le w \le 1$, $\boldsymbol{\phi}(\theta) = (\theta,\theta_2,1)$ where $-\infty < \theta_2 < \infty$ or $\boldsymbol{\phi}(\theta) = (\theta_1,\theta,0)$ where $-\infty < \theta_1 < \infty$. All of these possible mappings produce the same value of the auxiliary likelihood, which coincides with the value of the generative likelihood.

It is straightforward under the assumptions of Results 1–2 to show that pdBIL will target the true posterior as $n \to \infty$ if the true model is contained within the auxiliary model. When the generative model is a special case of the auxiliary model, using the notation of Cox and Wermuth (1990), the auxiliary and generative parameter can be written as $\boldsymbol{\phi} = (\boldsymbol{\theta_e},\boldsymbol{\gamma})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta_r},\boldsymbol{\gamma_0})$ where e and r denote 'extended' and 'reduced' respectively and $\boldsymbol{\gamma_0}$ is fixed. The proof of this result is straightforward. It involves demonstrating that $\boldsymbol{\phi}(\boldsymbol{\theta},\boldsymbol{x}_{1:n})$ is

consistent also for the parameter of the generative (reduced) model. Therefore when $n \to \infty$ the generative and auxiliary likelihoods will coincide.

This theoretical result cannot typically be realised in practice since a model which incorporates an intractable model as a special case is likely to also be intractable. However, it does suggest that the auxiliary model be chosen to be adequately flexible to give a good approximation of the generative likelihood for $\boldsymbol{\theta}$ values with positive prior support. In practice, our empirical evidence indicates that it is only necessary for the auxiliary model to mimic the behaviour of the generative likelihood for the values of $\boldsymbol{\theta}$ with non-negligible posterior support and for the auxiliary likelihood to be negligible in regions of the parameter space with little posterior support. If this is not the case it is likely that the pdBIL method will lead to poor approximations (as we demonstrate in Section 7.1).

If the binding function, $\boldsymbol{\phi}(\boldsymbol{\theta})$, were known, the pdBIL method would proceed straightforwardly. Since it will not be available in practice, it can be estimated indirectly through the simulated data $\boldsymbol{x}_{1:n}$. From the above, it is desirable for the pdBIL method if the auxiliary likelihood is as close as possible to the true likelihood. Gallant and McCulloch (2009) show, for a particular choice of the auxiliary model, that choosing the MLE for $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})$ minimises the Kullback-Leibler divergence between the generative and auxiliary likelihoods. Furthermore, this choice will often lead to Results 1–2 holding. Therefore we advocate the use of the MLE with this method. When the MLE of the auxiliary model is used, Cox (1961) provides an expression which defines $\boldsymbol{\phi}(\boldsymbol{\theta})$ (see equation (25) of Cox (1961)).

It remains to be seen what the target of $p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y})$ is relative to $p_A(\boldsymbol{\theta}|\boldsymbol{y})$. From an intuitive perspective, increasing n leads to a more precise determination of the mapping $\boldsymbol{\phi}(\boldsymbol{\theta})$, and thus should lead to a better approximation. A more theoretical argument is as follows. The approximations will coincide with each other provided $\mathsf{E}[p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})] = p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ (Andrieu and Roberts, 2009). Unfortunately, even if one uses an unbiased estimator for the auxiliary parameter, i.e. $\mathsf{E}[\boldsymbol{\phi}_{\boldsymbol{\theta},n}] = \boldsymbol{\phi}(\boldsymbol{\theta})$ for any value of n, this result will still rarely hold in general. The likelihood can be viewed here as a non-linear function of the auxiliary parameter estimate, and so $\mathsf{E}[p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})] \neq p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}))$ in general.Our empirical evidence (see Section 7) suggests that $p_{A,n}(\boldsymbol{\theta}|\boldsymbol{y})$ typically becomes less precise relative to $p_A(\boldsymbol{\theta}|\boldsymbol{y})$ as the likelihood estimate becomes more noisy, that is, when n is reduced. The message of Results 1–2 is that, provided the auxiliary model is suitably chosen, a better approximation can be anticipated by taking n as large as possible, which has the effect of reducing the bias of $p_A(\boldsymbol{y}|\boldsymbol{\phi}_{\boldsymbol{\theta},n})$.

4.1.2 MCMC pdBIL

As an example, MCMC can be used to sample from the pdBIL target (referred to here as MCMC pdBIL, see Gallant and McCulloch (2009)). This approach is presented in Algorithm 2.

As Results 1–2 suggest, the aim with pdBIL is to select n as large as possible. We demonstrate in the examples that it is computationally still feasible to consider values of n greater than one due to the improved statistical efficiency of MCMC pdBIL (and potentially other algorithms that implement pdBIL) when increasing n. Of course, the method will become computationally infeasible for very large n.

Algorithm 2 MCMC pdBIL algorithm (see also Gallant and McCulloch (2009)).

Set $\boldsymbol{\theta}^0$
Simulate $\boldsymbol{x}^*_{1:n} \sim p(\cdot|\boldsymbol{\theta}^0)$
Compute $\boldsymbol{\phi}^0 = \arg\max_{\boldsymbol{\phi}\in\Phi} p_A(\boldsymbol{x}^*_{1:n}|\boldsymbol{\phi})$
for $i = 1$ **to** T **do**
  Draw $\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{i-1})$
  Simulate $\boldsymbol{x}^*_{1:n} \sim p(\cdot|\boldsymbol{\theta}^*)$
  Compute $\boldsymbol{\phi}(\boldsymbol{x}^*_{1:n}) = \arg\max_{\boldsymbol{\phi}} p_A(\boldsymbol{x}^*_{1:n}|\boldsymbol{\phi})$
  Compute $r = \frac{p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{x}^*_{1:n}))p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)}{p_A(\boldsymbol{y}|\boldsymbol{\phi}^{i-1})p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})}$
  if uniform$(0,1) < r$ **then**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$
    $\boldsymbol{\phi}^i = \boldsymbol{\phi}(\boldsymbol{x}^*_{1:n})$
  else
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$
    $\boldsymbol{\phi}^i = \boldsymbol{\phi}^{i-1}$
  end if
end for

4.2 Parametric Bayesian Indirect Likelihood for a Summary Statistic (psBIL)

Wood (2010) proposes a method called synthetic likelihood when it is convenient to perform inference on the basis of a set of summary statistics rather than the full data. Considering Bayesian inference, the target distribution when the data have been reduced to a summary statistic is given by

$$p(\boldsymbol{\theta}|s(\boldsymbol{y})) \propto p(s(\boldsymbol{y})|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The major issue with this construction is that there is no analytical form for the likelihood function of the summary statistic, $p(s(\boldsymbol{y})|\boldsymbol{\theta})$. Wood (2010) overcomes this by applying, based on the terminology in this paper, a parametric auxiliary model for this probability distribution, $p_A(s(\boldsymbol{y})|\boldsymbol{\phi}(\boldsymbol{\theta}))$. In our framework, an approach that applies a parametric auxiliary model to form the likelihood of the summary statistic rather than the likelihood of the full data (as is presented in Section 4.1) is referred to here as psBIL (where 's' denotes summary statistic). Therefore the Bayesian version of Wood (2010) is a psBIL approach. For the auxiliary likelihood, $p_A(s(\boldsymbol{y})|\boldsymbol{\phi}(\boldsymbol{\theta}))$, Wood (2010) considers using the likelihood of a multivariate normal distribution, $N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ is the auxiliary parameter. As is the case with pdBIL, the binding function $\boldsymbol{\phi}(\boldsymbol{\theta})$ is rarely known but can be estimated via simulated data from the generative model for a particular value of $\boldsymbol{\theta}$. Using our notation, we obtain the following dataset from the true model of the summary statistic, $(s(\boldsymbol{x}_1), \ldots, s(\boldsymbol{x}_n))$. This represents n iid observations from $s(\cdot)|\boldsymbol{\theta}$. An advantage of selecting such a simple auxiliary

model is that the MLE has the analytic form

$$\boldsymbol{\mu}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} s(\boldsymbol{x}_i),$$
$$\boldsymbol{\Sigma}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} (s(\boldsymbol{x}_i) - \boldsymbol{\mu}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}))(s(\boldsymbol{x}_i) - \boldsymbol{\mu}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}))^T,$$

where the superscript T denotes transpose. The auxiliary likelihood used is then based on $N(\boldsymbol{\mu}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}))$. Our results indicate that the target distribution of this method will depend on n, and, if the auxiliary model for the summary statistic likelihood is reasonable, better approximations of $p(\boldsymbol{\theta}|s(\boldsymbol{y}))$ are likely to be obtained for large n.

5 ABC as a BIL method with Non-Parametric Auxiliary Model

An alternative and perhaps natural candidate for p_A is to use a kernel density estimate based on the samples $\boldsymbol{x}_{1:n}$. This corresponds to choosing $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) = \boldsymbol{x}_{1:n}$ and we define

$$p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})) = p_A(\boldsymbol{y}|\boldsymbol{x}_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} K_\epsilon(\rho(\boldsymbol{y}, \boldsymbol{x}_i)),$$

(Diggle and Gratton, 1984) where ϵ is the bandwidth parameter. We have then

$$\begin{aligned}
p_{A,n}(\boldsymbol{y}|\boldsymbol{\theta}) &= \int_{\mathsf{Y}^n} p_A(\boldsymbol{y}|\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})) \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\theta}) d\boldsymbol{x}_{1:n} \\
&= \int_{\mathsf{Y}^n} \frac{1}{n}\sum_{i=1}^{n} K_\epsilon(\rho(\boldsymbol{y}, \boldsymbol{x}_i)) \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\theta}) d\boldsymbol{x}_{1:n} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathsf{Y}^n} K_\epsilon(\rho(\boldsymbol{y}, \boldsymbol{x}_i)) \prod_{j=1}^{n} p(\boldsymbol{x}_j|\boldsymbol{\theta}) d\boldsymbol{x}_{1:n} \\
&= \int_{\mathsf{Y}} K_\epsilon(\rho(\boldsymbol{y}, \boldsymbol{x})) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \\
&\equiv p_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}),
\end{aligned}$$

and this is exactly the form of the standard ABC likelihood. In addition, n does not affect the likelihood (although it may help computationally in some algorithms) and ϵ controls the level of approximation. Here we see that this is an estimate of the ABC likelihood where the comparison is made between the full datasets. Here we obtain the npdBIL approach as presented in Figure 1 (where 'd' corresponds to full 'd'ata). Alternatively, a non-parametric density estimate of the auxiliary model of the summary statistic likelihood, $p_A(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\phi}(\boldsymbol{x}_{1:n}, \boldsymbol{\theta}))$, could be applied. Using a similar procedure to above, we obtain $p_\epsilon(s(\boldsymbol{y})|\boldsymbol{\theta})$. We refer to this in Figure 1 as npsBIL (npBIL based on a 's'ummary statistic). This is the approach adopted by Creel and Kristensen (2013), however only point estimation is considered. Unfortunately, Creel and Kristensen (2013) refer to their Bayesian estimator as BIL, however under our framework BIL is a more general class of methods. Of course, if the summary statistic is

derived from some parametric auxiliary model, then the ABC II class of method is recovered as an npsBIL method. The reader is again referred to Figure 1 to see the connection between these methods.

By selecting a parametric model for the auxiliary likelihood (pBIL), we can overcome the curse of dimensionality associated with the non-parametric aspect of ABC. Of course, finding a suitable parametric auxiliary model may be challenging in practice.

6 Comparison of ABC II and pdBIL

There are a few remarks to be made about the above results in relation to theoretical comparisons between ABC II and pdBIL.

Remark 2. *Under suitable conditions better approximations with pdBIL are obtained by increasing n. This is in stark contrast with ABC II, which cannot be trusted for $n > 1$.*

Remark 3. *In the case where the true model is a special case of the auxiliary model, the pdBIL method will be exact in the limit as $n \to \infty$. In contrast, in this ideal situation, ABC II still does not produce sufficient statistics (see the dimensionality argument in Section 3.5) and will not target the true posterior in the limit as $\epsilon \to 0$. Of course, finding an auxiliary model that satisfies this condition in practice will rarely be feasible.*

Remark 4. *Even if the auxiliary parameter estimate or score happen to be a sufficient statistic for the generative model, pdBIL still will not generally target the true posterior, as the auxiliary and generative likelihoods will still not match up. In this situation, the ABC II approaches will enjoy convergence to the true posterior as $\epsilon \to 0$ whilst pdBIL will not converge to the true posterior as $n \to \infty$. However, sufficient statistics are rarely achieved in practice.*

Remarks 3 and 4 demonstrate that pBII methods generally will not (and rarely will) target the true posterior distribution asymptotically. This is generally the case for other techniques in the literature for dealing with models that have intractable likelihood functions. There are some exceptions to this. For example, exact techniques are available for so-called doubly intractable models when perfect simulation from the generative model is possible (e.g. Murray et al. (2006)). Drovandi et al. (2013) show that exact Bayesian inferences can be produced for a certain class of models for low-count time series data using particle MCMC techniques (Andrieu et al., 2010). Despite not being exact, we demonstrate that pBII methods can produce quite good approximations in some applications.

The characteristics of a good auxiliary model differ between the ABC II and pdBIL methods. In the context of ABC II, we simply require a good summarisation of the data. That is, a low-dimensional and near-sufficient summary statistic. Therefore, we feel that it is useful if the auxiliary model in this context provides a good fit to the data and is parsimonious, so that the essential features of the data are described well and as succinctly as possible. This is completely independent of the process for selecting a generative model. Therefore the same auxiliary model will be used regardless of which generative model is fit to the data. For pdBIL, we require a flexible auxiliary model that can mimic the behaviour of the generative model for different values of $\boldsymbol{\theta}$ within the posterior support. Here, it is not necessary for the auxiliary

model to provide a good fit to the data considering the fact that the generative model being proposed might be mis-specified. The auxiliary model chosen for pdBIL may alter depending on the generative model being proposed. In our examples, the generative model is either known or provides a good fit to the data. In such cases, it would not be uncommon to choose the same auxiliary model for the ABC II and pdBIL methods.

The conditions required for pdBIL to produce exact results are very strong and finding an auxiliary model that is sufficiently flexible so that the auxiliary likelihood can mimic the generative likelihood could be difficult in practice. In some applications, an auxiliary model that is a simplified version of the generative model may be specified where the parameter of each model has the same interpretation. For example, the auxiliary model for a continuous time Markov jump process may be its corresponding linear noise approximation. In such situations, the pdBIL method is unlikely to perform well whilst it remains possible that such an approximate model could produce useful summary statistics for ABC even though the auxiliary model would not fit the data well. Jiang and Turnbull (2004) show that II can work well in the classical framework when the auxiliary model is a simplified version of the generative model. Further research is required in the Bayesian setting.

An additional advantage of the ABC II approach over pdBIL is the extra flexibility of being able to accommodate additional summary statistics that do not involve an auxiliary model, since this method belongs in the more general npsBIL class (see Creel and Kristensen (2013) and Wood (2010) for examples where the summary statistic is a combination of auxiliary parameter estimates and other summary statistics). Jiang and Turnbull (2004) and Heggland and Frigessi (2004) consider II applications in a classical framework where the comparison of observed and simulated data is made on the basis of both an auxiliary model and supplementary summary statistics.

7 Examples

7.1 Toy Example 1

In this example we consider a simple model so that exact Bayesian inferences are trivially obtained. Our intention here is to investigate the theoretical considerations in Section 4. In particular we show that, when the auxiliary model is reasonable, BIL produces better approximations as the size of simulated datasets goes beyond that of the observed data and as a useful by-product increases the acceptance probability of the MCMC moves. We also demonstrate empirically that unfortunately ABC approaches (including those using II to obtain summary statistics) do not possess this same desirable property as n is increased. Additionally, we investigate the output of pdBIL when the auxiliary model is poorly chosen.

Here the data are $N = 100$ independent draws from a Poisson distribution with a mean of $\lambda = 30$, $\boldsymbol{y} = (y_1, \ldots, y_{100}) \overset{\text{iid}}{\sim} Po(30)$. The prior is $\lambda \sim \Gamma(\alpha, \beta)$ (where $\alpha = 30$ and $\beta = 1$), which results in a $\lambda|\boldsymbol{y} \sim \Gamma(30 + \sum_{i=1}^{100} y_i, 101)$ posterior. For such a relatively large value for the mean of the Poisson distribution, a normal approximation with mean, μ, and variance, τ, will be reasonable. We use this normal distribution as the auxiliary model. Here the auxiliary likelihood, MLE and score are trivial to compute. The Anderson-Darling test for normality produced a p-value of about 0.576, which indicates no evidence against the assumption that

the normal auxiliary model provides a good description of the data.

The summary statistic based on this auxiliary model includes the sample mean, $\bar{y}$, which is a sufficient statistic for the generative model. Thus, the ABC II approaches can be expected to produce essentially exact inferences (excluding Monte Carlo error) as long as the ABC tolerance is low enough. As demonstrated in Figure 2, this is the case. Such sufficiency is not usually achieved in practice. However, it can be seen that the ABC posterior is grossly over-precise when the simulated datasets size is increased to 1000 (i.e. $n = 10$).

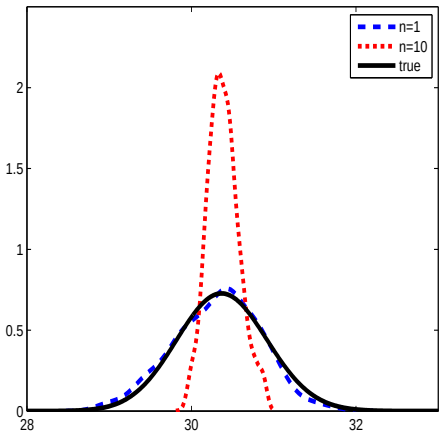
In the limit as $n \to \infty$ the pdBIL method boils down to a $N(\lambda, \lambda)$ distribution approximating a $Po(\lambda)$ distribution. The central limit theorem states that the normal approximation improves as λ increases. Since $\lambda = 30$, pdBIL can never target the true posterior. The pdBIL target distribution (for $n \to \infty$) is proportional to $\lambda^{\alpha - N/2 - 1} \exp(-(\beta + N/2)\lambda) \exp(-\sum_{i=1}^{N} y_i^2/(2\lambda))$ while the true posterior is proportional to $\lambda^{\alpha + \sum_{i=1}^{N} y_i - 1} \exp(-(\beta + N)\lambda)$. Figure 2(d) demonstrates a small amount of bias for the pdBIL method (this is an illustration of Remark 4).

Figure 2(d) presents the results for the pdBIL method based on simulated dataset sizes of $n = 1$ and $n = 10$ (results for $n = 100$ and $n = 1000$ are even closer to the true posterior but are not shown on the figure). It is evident from the figure that a more precise posterior is achieved when using larger simulated datasets, without necessarily over-shooting the true posterior. Additionally, there was an increase in the MCMC acceptance rate as n increased. For the n values investigated here, the acceptance rates were roughly 46%, 67%, 72% and 73% for increasing n. These acceptance rates are very high, especially relative to ABC algorithms which generally suffer from quite low acceptance probabilities. This example demonstrates that better inferences using pdBIL can be obtained by increasing the size of the simulated dataset beyond that of the observed. Unfortunately ABC inferences that use a simulated data size larger than that of the observed data cannot be trusted in the same way (see Remark 2).

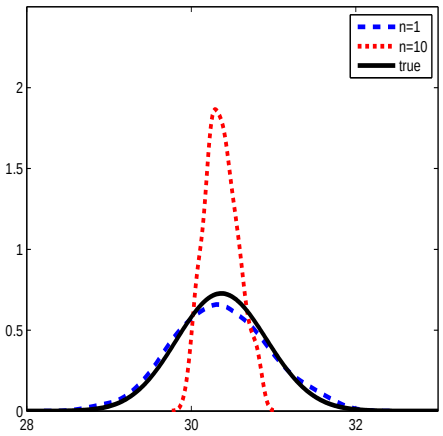
The reason for improved inferences from pdBIL as n is increased is apparent from Figure 3. Here it can be seen from increasing n the log-likelihoods of the generative and auxiliary models are becoming more correlated with the slope of the relationship becoming approximately one.

Figure 4(a) shows the true $Po(\lambda)$ and auxiliary $N(\lambda, \lambda)$ log-likelihood values for λ within the 99% highest prior density region. The vertical lines indicate the bounds of a 99% credible interval based on the true posterior. It can be seen that the auxiliary log-likelihood is a poor approximation to the true likelihood in regions with negligible posterior support. This is even the case for larger λ values where it would be expected that a normal approximation would be more appropriate. However, the normal approximation will perform relatively poorly in the tails of the distribution. It is evident that the auxiliary likelihood acts as a useful replacement likelihood in the region of high posterior support (see Figure 4(c)), and this is enough to result in a good approximation of the true posterior for large n.

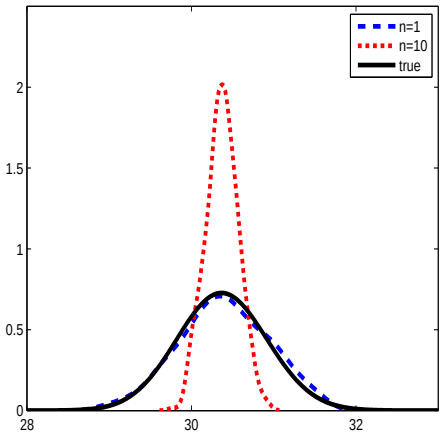
Finally we investigate the output from pdBIL when the auxiliary model is chosen poorly. Figure 4(c) shows the results for when the auxiliary model is $N(\mu, \tau_0)$, where τ_0 is fixed. Here the pdBIL posterior as $n \to \infty$ is proportional to $\lambda^{\alpha - 1} \exp(-\lambda(\beta + \sum_{i=1}^{n} y_i/\tau_0)) \exp(-n\lambda^2/(2\tau_0))$. Here we consider $\tau_0 = 49$ (overdispersed) and $\tau_0 = 16$ (underdispersed). The results show over-precise and conservative results in the underdispersed and overdispersed case, respectively. The underdispersed and overdispersed auxiliary models have thinner and fatter tails, respectively, than the likelihood of the generative model in the parameter space well sup-



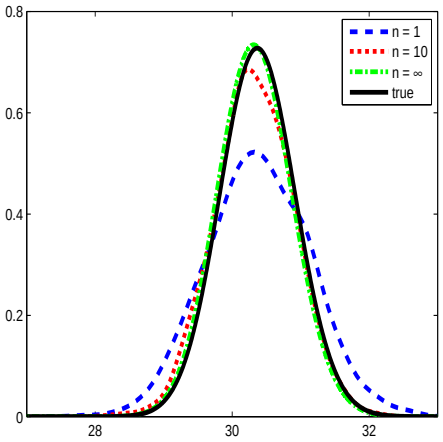
(a) ABC IP



(b) ABC IS

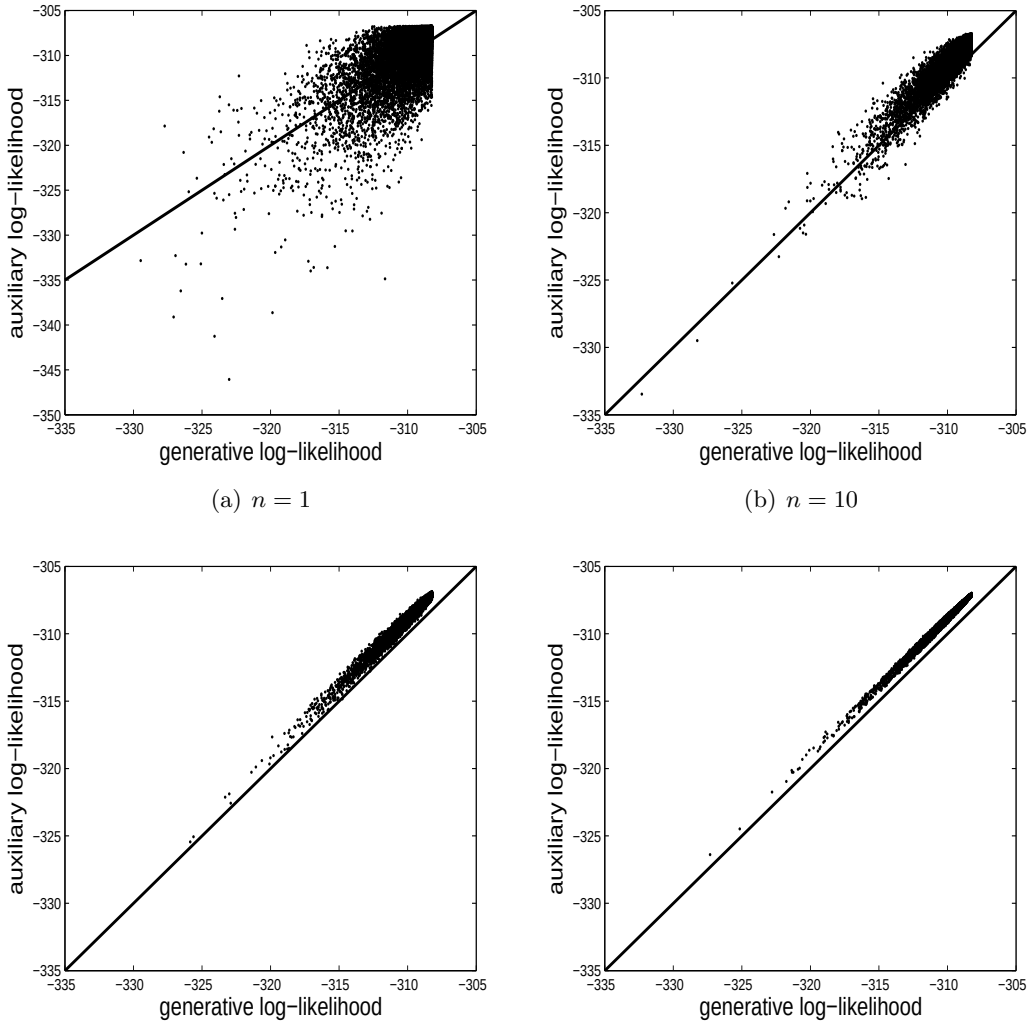


(c) ABC IL



(d) BIL

Figure 2: Posterior distributions for λ, the parameter of the Poisson example for when (a) ABC IP, (b) ABC IS, (c) ABC IL and (d) pdBIL is applied (on-line figure in colour).



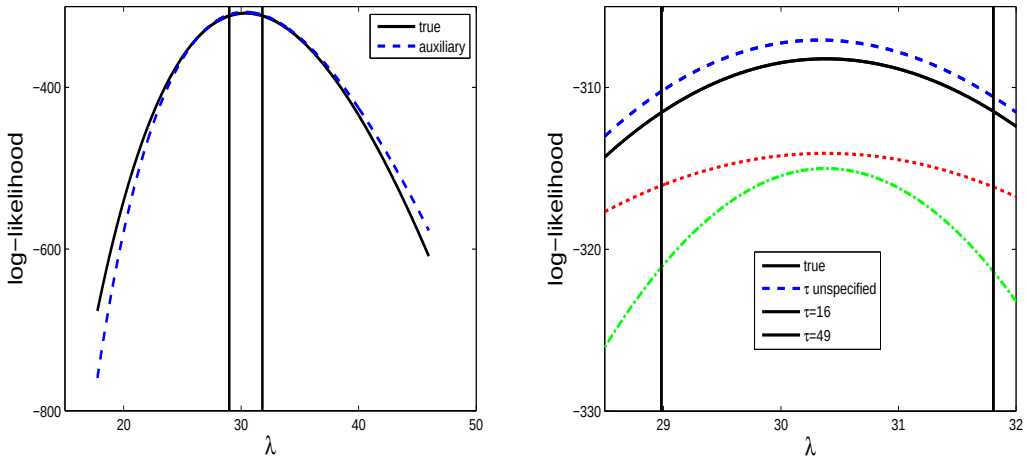
(a) $n = 1$

(b) $n = 10$

(c) $n = 100$

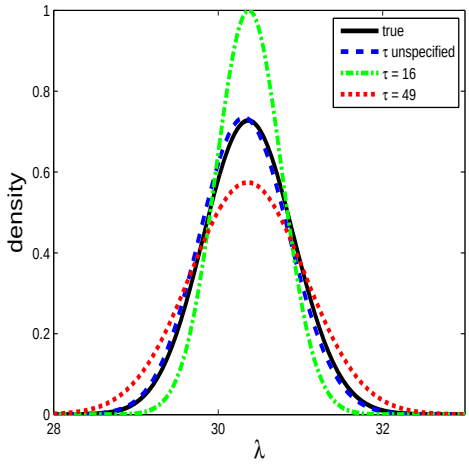
(d) $n = 1000$

Figure 3: Comparison of the generative and auxiliary log-likelihoods for the toy example calculated during the MCMC pdBIL algorithm with different values of n ((a) $n = 1$, (b) $n = 10$, (c) $n = 100$, (d) $n = 1000$).



(a)

(b)



(c)

Figure 4: Investigating different auxiliary models for the first toy example. (a) Comparison of the true and auxiliary log-likelihoods values for λ values within the 99% highest prior density region. (b) Comparison of true and auxiliary log-likelihoods for different choices of the auxiliary model. The vertical lines in (a) and (b) indicate the bounds of a 99% credible interval based on the true posterior. (c) Comparison of the posterior distributions for the three different auxiliary models (on-line figure in colour).

ported by the posterior distribution (see Figure 4). In these cases the auxiliary model is not providing a useful replacement likelihood. Just by chance ABC II is still exact here as $\epsilon \to 0$ since the parameter estimate for μ is a sufficient statistic for λ.

7.2 Toy Example 2

In this example the true model is a t-distribution with mean, standard deviation and degrees of freedom of μ, σ and 1 respectively. The auxiliary model is a more general t-distribution with degrees of freedom ν. In this case, the pdBIL method is exact in the limit as $n \to \infty$ as the true model is incorporated within the auxiliary model. Unfortunately, ABC II does not produce a sufficient statistic as the summary statistic will be of dimension three whilst it is known for this model the minimal sufficient statistic consists of all the order statistics. This example demonstrates Remark 3.

7.3 g-and-k Example

7.3.1 Models and Data

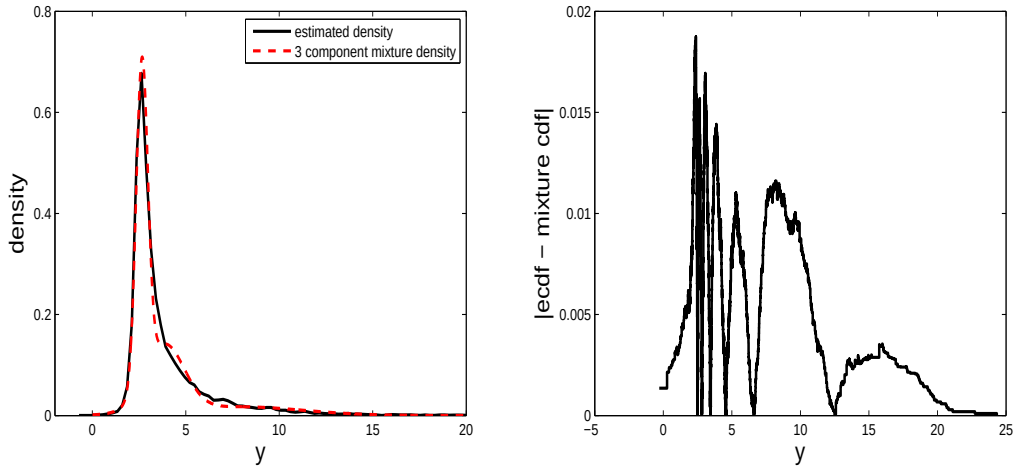
Quantile distributions (or functions) represent a class of distributions that are defined in terms of their quantile function. Such functions can be formulated to create more flexible distributions than other standard distributions. In this example the focus is on the g-and-k distribution described in, for example, Rayner and MacGillivray (2002) (the reader is also referred to the references therein). This quantile function, which can also be interpreted as a transformation of a standard normal random variate, has the following form

$$Q(z(p); \boldsymbol{\theta}) = a + b \left(1 + c\frac{1 - \exp(-gz(p))}{1 + \exp(-gz(p))}\right) (1 + z(p)^2)^k z(p). \tag{2}$$

Here p denotes the quantile of interest while $z(p)$ represents the quantile function of the standard normal distribution. The model parameter is $\boldsymbol{\theta} = (a, b, c, g, k)$, though common practice is to fix c at 0.8, which we do here (see Rayner and MacGillivray (2002) for a justification). The likelihood function can be computed numerically, although is more expensive than model simulation which is naturally implemented for quantile distributions via the inversion method. Full likelihood-based inference is more expensive than the simulation-based approaches for the relatively large dataset considered here.

The observed dataset consists of 10000 independent draws from the g-and-k distribution with $a = 3$, $b = 1$, $c = 0.8$, $g = 2$ and $k = 0.5$ (same as considered in Drovandi and Pettitt (2011)). A non-parametric estimate of the probability density function based on these samples is shown in Figure 5. The data exhibit significant skewness and kurtosis.

We use a three component normal mixture model with 8 parameters as the auxiliary model. A mixture model is a suitable choice for an auxiliary distribution since it can be made arbitrarily flexible whilst maintaining a tractable likelihood function. Therefore auxiliary MLEs are computationally easy to obtain (here we use the Expectation-Maximisation algorithm) and the subsequent likelihood can be evaluated cheaply. On the other hand, mixture models can he highly irregular and the MLE is not consistent in general. The invariance of the likelihood



(a) (b)

Figure 5: (a) Non-parametric estimate of the density function based on a dataset simulated from a g-and-k distribution together with the density of a three component mixture of normals estimated from the data. (b) Comparison of the theoretical cdf of the three component mixture model with the empirical cdf of the data (on-line figure in colour).

to a re-labelling of the components causes an immediate issue for ABC IP, which requires a unique auxiliary parameter estimate. This could be remedied by applying an identifiability constraint to the mixture model (for example, ordering the means). However this reduces the flexibility of the auxiliary model, increases the difficulty of implementation and does not necessarily solve the invariance issue. For these reasons, we do not consider ABC IP suitable in this example. Since pdBIL and ABC IL use the likelihood of the auxiliary model, it can potentially accommodate the invariant property of the mixture model. At least the labelling issue can be safely ignored with these approaches. However, the mixture model can result in other numerical issues such as infinite likelihoods. This would create serious issues for methods that use the auxiliary likelihood (the auxiliary likelihood would not be unique). From investigations on the dataset here, it appears that the likelihood is well behaved and that the modes in the likelihood correspond only to re-labelling of components. Therefore we proceed with ABC IL and pdBIL with caution. The ABC IS method, based on the score vector, appeared to not have any difficulties accommodating the auxiliary mixture model.

From Figure 5 it appears that the auxiliary model provides a reasonable fit to the data. Figures 5(a) and 5(b) it can be seen that there is a correspondence between both the densities and the cumulative distribution functions of the mixture model and the data. However, we performed a hypothesis test to assess the goodness-of-fit of the three component mixture model with a parameter given by the MLE. The test-statistic was the Kolmogorov-Smirnov statistic that computes the maximum absolute difference between the theoretical and empirical cdfs. To avoid any distributional assumption about this test-statistic, we simulated 10000 values of this statistic under the assumption that mixture model is correct. We found that the observed test-statistic was exceeded only 0.25% of the time, indicating strong evidence against the mixture providing a good fit to the data. Figure 5(b) shows the differences between the empirical

and theoretical cdfs. However, from Figure 5(a) it is evident that the mixture model can explain several features of the true model, and since the dataset size is large there is a high probability of detecting a difference. Our results below show that we are able to obtain quite accurate posterior distributions with the pBII methods despite the lack of fit suggested by the hypothesis test. In Appendix B, we present results for using a four component mixture model. Unfortunately we found this was substantially more expensive to apply and resulted in some numerical problems.

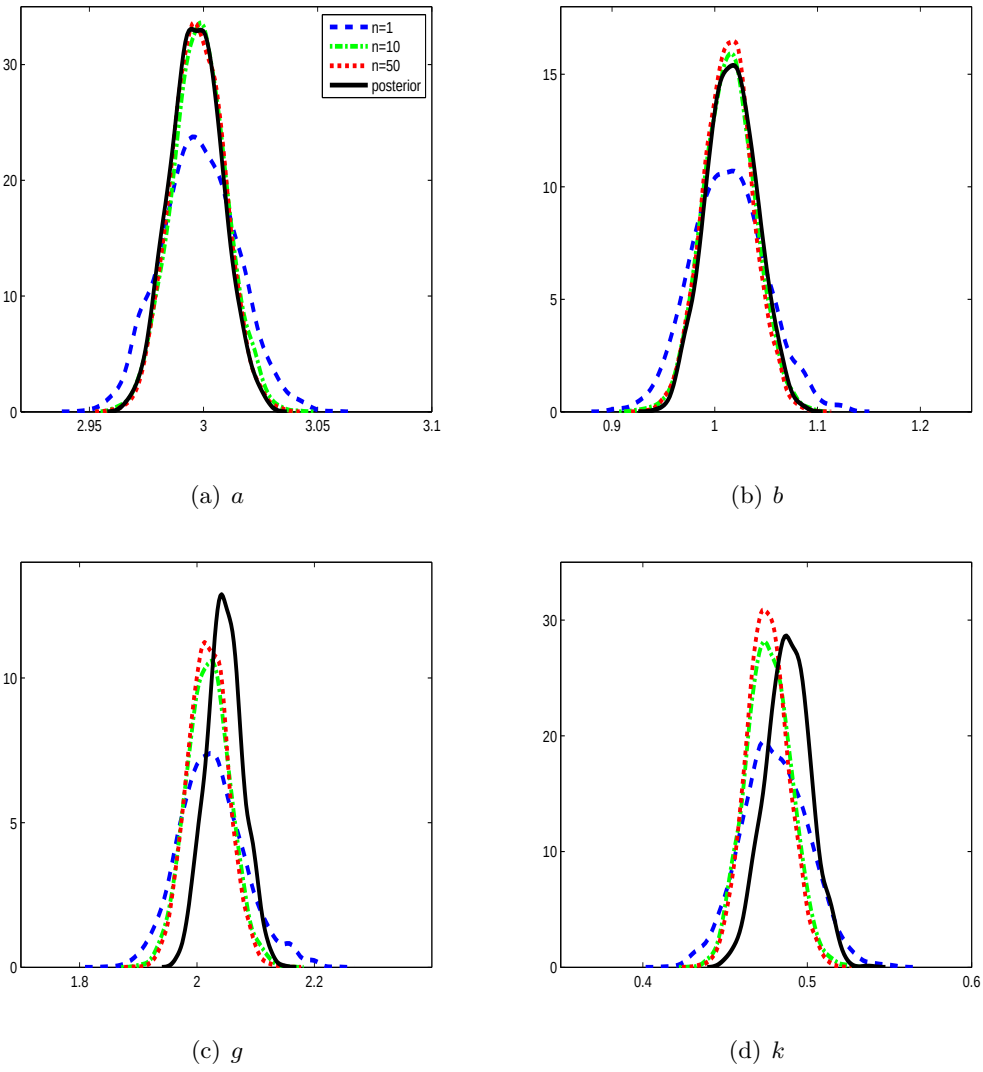
7.3.2 Results

The proposal distribution in the MCMC for the pdBIL algorithm was guided using the results in Drovandi and Pettitt (2011), who analysed the same data via a traditional ABC approach that used robust measures of location, scale, skewness and kurtosis as the summary statistics.

pdBIL was run using n values of 1, 2, 4, 10, 20 and 50 for a number of iterations given by 1 million, 500000, 500000, 200000, 100000 and 75000 respectively. The acceptance probabilities obtained were about 2.8%, 5%, 7.1%, 13.1%, 18.5% and 20.8% respectively. The average effective sample size (ESS, averaged over the four parameters) divided by the computing time (in hours) was roughly 63, 127, 106, 124, 70 and 41 respectively. This demonstrates how pdBIL is still feasible as n increases to a certain point. However for very large n the computation becomes unmanageable.

Figure 6 shows the results for $n = 1$, $n = 10$ and $n = 50$ (the results for $n = 20$ and $n = 50$ were quite similar). A very time consuming exact MCMC algorithm was run for 10000 iterations to obtain a gold-standard (producing an average ESS per hour of 6). The results show an increase in precision of the pdBIL results as n increases. The results for a and b are very accurate, while the pdBIL posteriors for g and k show some bias (also with a loss of precision for g).

ABC IL was run for 1 million iterations with a tolerance tuned to achieve an acceptance rate of about 1%. Due to the ABC IS method being so much faster than the other pBII approaches, we aimed for a relatively lower ABC tolerance and ran the algorithm for more iterations. More specifically, 7 million iterations were used and the ABC tolerance chosen resulted in an acceptance rate of 0.3%. We also applied a regression adjustment to the (appropriately thinned) ABC IS samples using the approach of Beaumont et al. (2002) in an attempt to eliminate the effect of the ABC tolerance. The regression adjustment could not be applied to ABC IL since the auxiliary summary statistics are multi-modal owing to the invariant nature of the auxiliary likelihood. The ABC IS gave slightly better results than ABC IL (results not shown in Figure 7), which could be due to the ability of ABC IS getting to a lower ABC tolerance due to the increase in efficiency. ABC IS produced an average ESS per hour of 90 while the corresponding number was 30 for ABC IL, showing that the ABC IS method required less time to produce a better approximation. Furthermore, the regression adjustment gave some improvement to the ABC IS results. We compared the pBII approaches with the ABC results of Drovandi and Pettitt (2011). It should be noted that we applied a local regression adjustment to the ABC results here as we found some improvement for the parameters a and g (results were very similar for b and k relative to those obtained in Drovandi and Pettitt (2011)). The results are shown in Figure 7. Overall, the pBII results present a marked



(a) a (b) b

(c) g (d) k

Figure 6: Posterior distributions for the parameters of the g-and-k model based on the pdBIL approach with $n = 1$ (blue dashed), $n = 10$ (green dot-dash) and $n = 50$ (red dot). Also shown are results based on using the true likelihood (black solid) (on-line figure in colour).

improvement over the ABC analysis of Drovandi and Pettitt (2011). In particular, the ABC IS with regression adjustment probably produced the best results overall, with g seemingly the most difficult parameter to estimate. We also found that the ABC IS method with regression adjustment produced the best results when a four component auxiliary mixture model was used.

7.4 Macroparasite Example

7.4.1 Models and Data

Drovandi et al. (2011) developed an ABC IP approach to estimate the parameters of a stochastic model of macroparasite population evolution developed by Riley et al. (2003) (see also Michael et al. (1998)). Data was collected independently on 212 cats, who were initially injected with a certain number of juvenile *Brugia pahangi* parasites. Some time after each cat was sacrificed, the number of parasites that had reached maturity were counted and recorded (see Denham et al. (1972)). Drovandi et al. (2011) discovered that a beta-binomial model provided a good description of the data. Below provides a brief review of the generative and auxiliary models, with the reader referred to Drovandi et al. (2011) for more details.

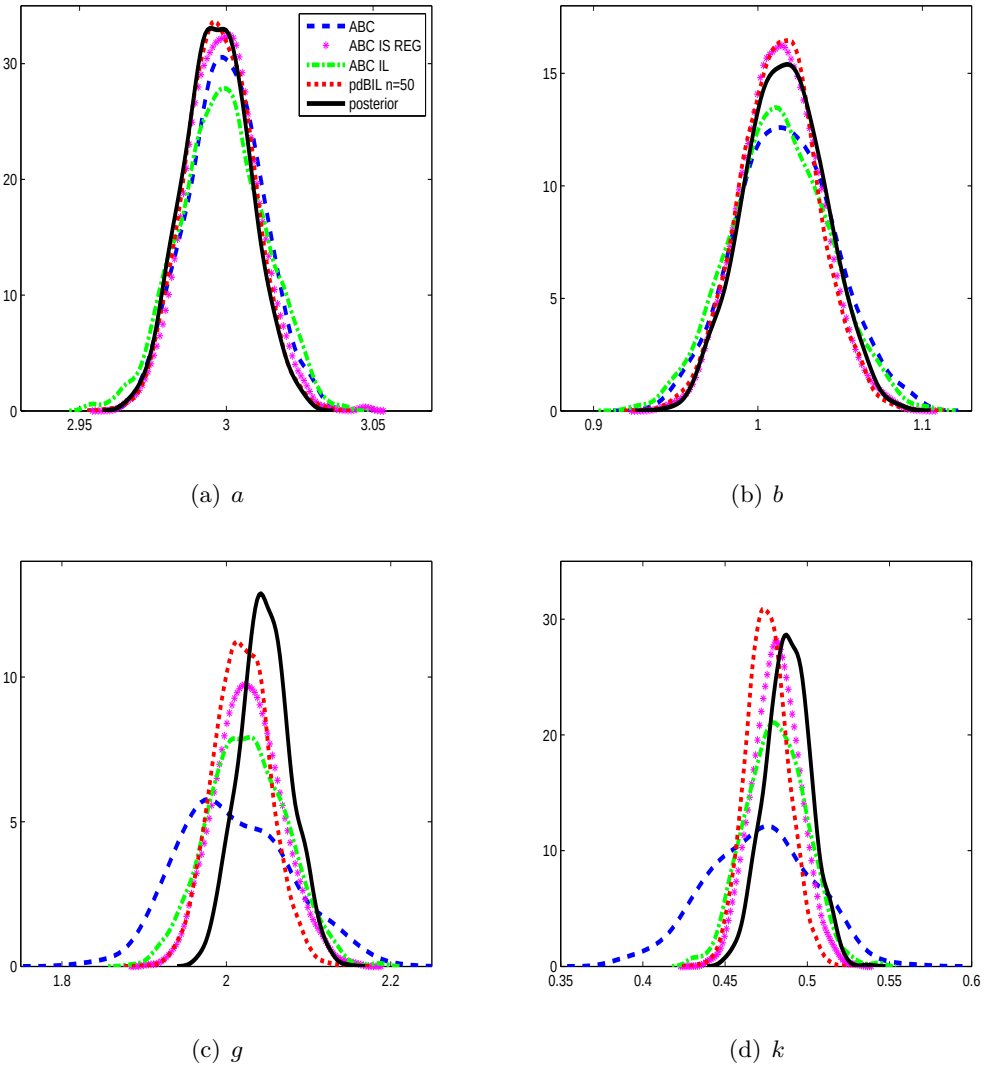
At time t any host is described by three random variables $\{M(t),\, L(t),\, I(t)\}$, where $M(t)$ is the number of mature parasites, $L(t)$ is the number of larvae and $I(t)$ is a discrete version of the host's immunity.

It is assumed that each larva matures at a constant rate of γ per day. Larvae die at a rate $\mu_L + \beta I(t)$ per larva where μ_L represents the rate at which natural death of larvae occurs and β is a rate parameter that describes additional death of larvae due to the immune response of the host. The acquisition of immunity is assumed to be dependent only on the number of larvae and occurs at rate $\nu L(t)$, and a host loses immunity at a rate μ_I per unit of immunity. Mature parasites die at a rate of μ_M adults per day. Parameters γ and μ_M have been previously estimated at 0.04 (Suswillo et al., 1982) and 0.0015 (Michael et al., 1998) respectively.

The data were modelled via a continuous time discrete trivariate Markov process. Given current values of the states at time t, $M(t) = i, L(t) = j, M(t) = k$, and a small time increment Δ_t the transition probabilities at time $t + \Delta_t$ are given by

$$\begin{aligned} p(i+1, j-1, k) &= \gamma j \Delta_t + o(\Delta_t), \\ p(i, j-1, k) &= (\mu_L + \beta k) j \Delta_t + o(\Delta_t), \\ p(i-1, j, k) &= \mu_M i \Delta_t + o(\Delta_t), \\ p(i, j, k+1) &= \nu j \Delta_t + o(\Delta_t), \\ p(i, j, k-1) &= \mu_I k \Delta_t + o(\Delta_t), \end{aligned} \tag{3}$$

and the probability of remaining in the same state is one minus the sum of the above probabilities. Only the final mature count is observed whilst the immunity and larvae counts are unobserved throughout the process. Moreover, the immune response variable $I(t)$ is unbounded. Data generative likelihood-based approaches appear infeasible due to computational issues (see Drovandi et al. (2011)). Simulation is straightforward via the algorithm of Gillespie (1977). The prior distributions are: $\nu \sim U(0,1)$, $\mu_I \sim U(0,2)$, $\mu_L \sim U(0,1)$ and $\beta \sim U(0,2)$.



(a) a (b) b

(c) g (d) k

Figure 7: Posterior distributions for the parameters of the g-and-k model based on the ABC approach of Drovandi and Pettitt (2011) (blue dash), ABC IS with regression adjustment (purple star), ABC IL (green dot-dash) and pdBIL with $n = 50$ (red dot) and results from using the true likelihood (black solid) (on-line figure in colour).

Here we denote the observed data as $\boldsymbol{y} = (m_1, \ldots, m_{212})$ where m_i is the mature count for the ith host. Covariates for the ith host are given by l_i (initial larvae count) and t_i (sacrifice time).

For the auxiliary model, Drovandi et al. (2011) capture the overdispersion via a beta-binomial generalised linear model and take into account the effect that t_i and l_i have on m_i. Denote α_i and β_i as the beta-binomial parameters for the ith host. More specifically, the ith observation has the following probability distribution

$$p(m_i|\alpha_i, \beta_i) = \binom{l_i}{m_i} \frac{B(m_i + \alpha_i, l_i - m_i + \beta_i)}{B(\alpha_i, \beta_i)}, \tag{4}$$

where $B(\cdot|\cdot)$ denotes the beta function. Consider a re-parameterisation in terms of a proportion, $p_i = \alpha_i/(\alpha_i + \beta_i)$, and over-dispersion, $\xi_i = 1/(\alpha_i + \beta_i)$, parameter. The auxiliary model relates these parameters to the covariates via

$$\begin{aligned} \text{logit}(p_i) &= f_p(t_i, l_i), \\ \log(\xi_i) &= f_\xi(t_i, l_i), \end{aligned}$$

where

$$f_\xi(t_i, l_i) = f_\xi(l_i) = \begin{cases} \eta_{100}, & \text{if } l_i \leq 100 \\ \eta_{200}, & \text{if } l_i > 100 \end{cases},$$

and

$$f_p(t_i, l_i) = f_p(t_i) = \beta_0 + \beta_1(\log(t_i) - \overline{\log(t)}) + \beta_2(\log(t_i) - \overline{\log(t)})^2.$$

Hence the auxiliary model has the parameter $\boldsymbol{\phi} = (\beta_0, \beta_1, \beta_2, \eta_{100}, \eta_{200})$ while the generative model has the parameter $\boldsymbol{\theta} = (\nu, \mu_I, \mu_L, \beta)$.

Using the approach outlined in Appendix C, we obtained goodness-of-test p-values of 0.37 and 0.47, indicating no evidence against the beta-binomial model providing a good description of the data. Drovandi et al. (2011) use the AIC to select this auxiliary model over competing auxiliary models.

7.4.2 Results for Simulated Data

For validation of the pBII methods for this example, data was simulated using the same experimental design as the observed data based on the parameter configuration estimated by Riley et al. (2003); $\nu = 0.00084$, $\mu_I = 0.31$, $\mu_L = 0.0011$ and $\beta = 1.1$. We found that the pBII methods were able to recover the parameters ν and μ_L well, μ_I was determined less precisely and β was not recovered. The data are not particularly informative about μ_I and β (see Drovandi et al. (2011) for more discussion). The ABC IS gave the most precise posterior distributions for ν and μ_L out of the pBII methods. For full details on the analysis of this simulated data, see Appendix C.

7.4.3 Results for Real Data

Here we used the ABC IP results of Drovandi and Pettitt (2011) to form an MCMC proposal distribution. The pdBIL method with $n = 1$, $n = 20$ and $n = 50$ was run for 1 million, 100000 and 50000 iterations, respectively. Acceptance probabilities of roughly 1.4%, 23.5% and 28.2%, respectively, were obtained. The average ESS per hour was 37, 79 and 58, respectively. The massive increase in acceptance probability allowed us to use fewer iterations. The results are shown in Figure 8. The figures suggest that we are not able to gain any additional information from the data for the parameter ν from the pdBIL approach by increasing n. However, an increase in precision is obtained for μ_L as n is increased. The posteriors are shifted slightly for the other two parameters, however are still largely uninformative, although the posterior for μ_I for large n may indicate some preference for smaller values of μ_I.

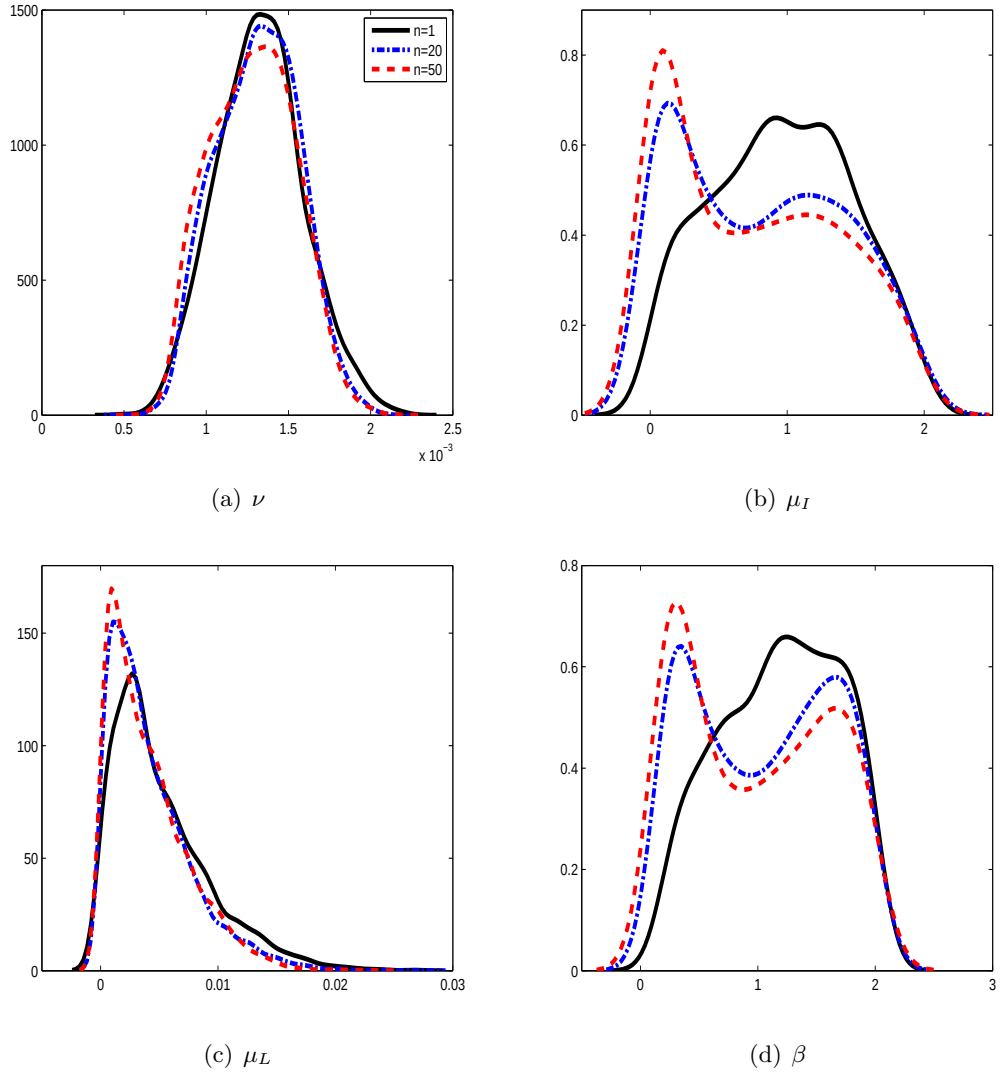
We now compare the results of pdBIL with ABC. ABC IP and ABC IL MCMC algorithms were all run for 1 million iterations. The ABC IP and ABC IL tolerances were chosen so that the acceptance rate was about 1.5%. pdBIL had a similar acceptance probability. Due to the increased computational efficiency of ABC IS, we ran this algorithm for 20 million iterations and tuned the tolerance to obtain an acceptance rate of about 0.1%. ABC IP and ABC IL used about 15 hours of computing time while ABC IS only required 11 hours even though 20 times more iterations were run.

The estimated posterior densities (after appropriate thinning) for the different approaches are presented in Figure 9. In general the data are not informative about the μ_I and β parameters, so we turn our focus to the parameters ν and μ_L. We note that it is difficult to compare the approximations without having available a gold standard. It can be seen that pdBIL produces the most precise inferences for the parameters ν and μ_L. Despite being able to reduce the ABC tolerance, the ABC IS method appears to be the least precise. This is in contrast to the results for the simulated data in Appendix C, where ABC IS produced the most precise results.

Regression adjustment was also applied to the ABC II methods in an attempt to reduce the effect of the ABC tolerance. These adjustments were applied individually to $-\log(\nu)$ and $\sqrt{\mu_L}$ (see Appendix C) and the results are shown in Figure 10. The regression adjustment does increase the precision of the ABC II posteriors. The regression adjustment appears to shift the modes of the ABC II results slightly for ν. For μ_L, the regression adjustment brings the ABC II results closer to that obtained by pdBIL for $n = 50$.

8 Discussion

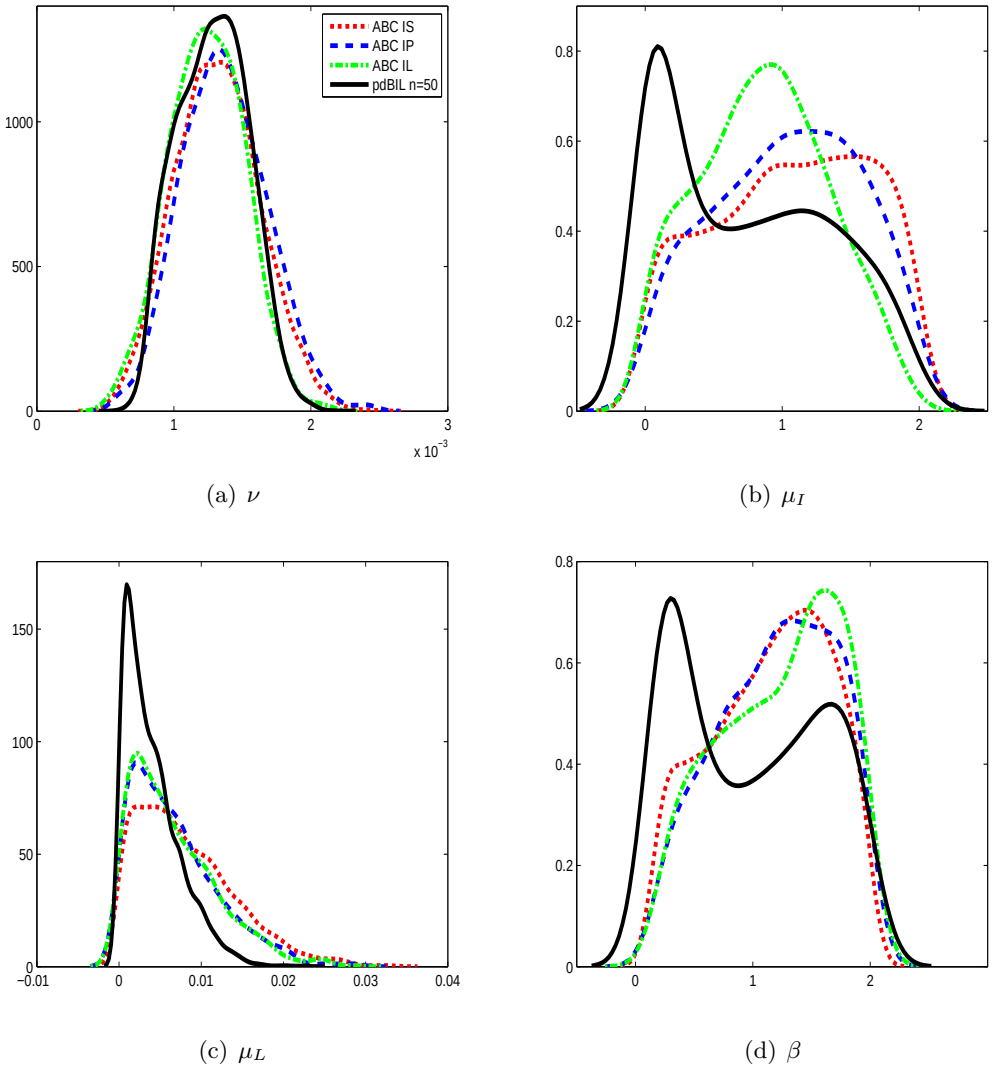
This paper has provided an extensive comparison of pBII methods, from theoretical, practical and empirical perspectives. We discovered that the pdBIL method of Gallant and McCulloch (2009) is fundamentally different to ABC II approaches developed in the literature. More specifically, we showed that pdBIL can produce better approximations by increasing the size of the simulated datasets as long as the auxiliary model can fit data drawn from the generative model for a variety of $\boldsymbol{\theta}$ values. In contrast, ABC methods (including those that use II to form the summary statistic) should simulate datasets the same size as the observed. The pdBIL method has the additional advantage of not having to determine an appropriate ABC



(a) ν (b) μ_I

(c) μ_L (d) β

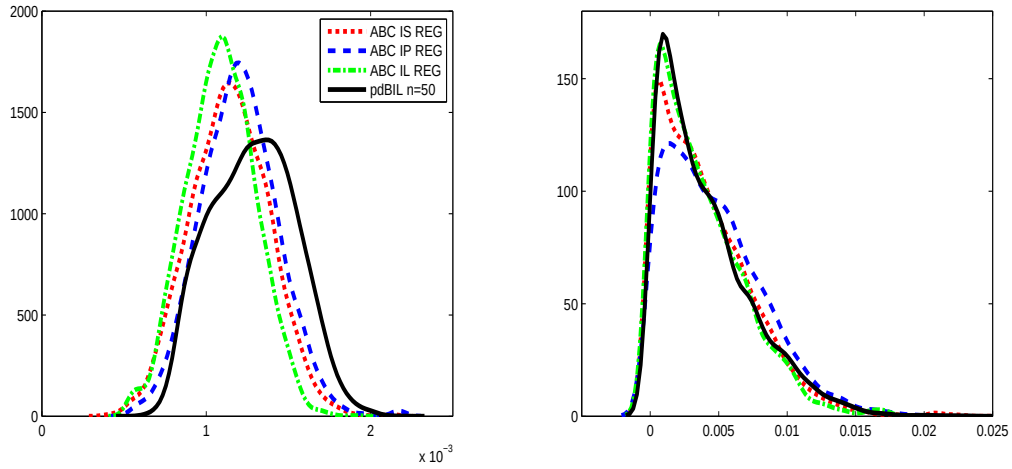
Figure 8: Posterior distributions for the parameters ((a) ν, (b) μ_I, (c) μ_L, (d) β) of the macroparasite model based on an pdBIL approach with $n = 1$ (black, solid), $n = 20$ (green, dashed) and $n = 50$ (red, dotted) (on-line figure in colour).



(a) ν (b) μ_I

(c) μ_L (d) β

Figure 9: Posterior distributions for the parameters ((a) ν, (b) μ_I, (c) μ_L, (d) β) of the macroparasite model based on ABC IP (blue dash), ABC IS (red dot), ABC IL (green, dot-dash) and pdBIL (black solid) (on-line figure in colour).



(a) ν (b) μ_L

Figure 10: Posterior distributions for the parameters ((a) ν and (b) μ_L) of the macroparasite model based on ABC IP (blue dash), ABC IS (red dot), ABC IL (green dot-dash) and pdBIL (black solid). Regression adjustment has been performed on the output of the ABC II approaches (on-line figure in colour).

tolerance. Furthermore, we found that increasing the size of the simulated dataset beyond that of the observed does not necessarily make computation infeasible due to the increase in statistical efficiency. However, it is of interest to determine the size of the dataset upon which negligible improvement will be obtained. This requires further research.

We have also established that BIL is a rather flexible framework since the synthetic likelihood approach of Wood (2010) is a pBIL method that applies a parametric auxiliary likelihood to the summary statistic likelihood while ABC can be recovered by selecting a specific non-parametric auxiliary model. Our focus in this paper has been on the pBIL method where an auxiliary model is proposed for the full data likelihood. However, the ideas in this paper can be carried over to when the auxiliary model is applied to a summary statistic likelihood as in Wood (2010).

For the pdBIL method to have some chance of a good approximation to the true posterior for the specified generative model, it is important that the auxiliary model is able to fit well data simulated from the generative model for parameter values within non-negligible posterior regions, at least in the majority of simulations. It would be possible to perform a goodness-of-fit test on the auxiliary model for every dataset generated from the proposed model during the MCMC pdBIL algorithm in order to assess the usefulness of the auxiliary model in the context of the pdBIL method. This is the subject of further research.

In this paper we have not addressed the issue of which ABC II method provides the best approximation. ABC IS is much faster (when the auxiliary score vector is analytic) and requires only weak assumptions, but did not always outperform the other ABC II methods in the examples considered in this paper. The ABC IP and ABC IL methods differ only in their discrepancy function and it is not clear if one discrepancy function dominates the other

across applications. Furthermore, it remains unknown if the auxiliary parameter estimate or auxiliary score carries the most information in the observed data. It could be that the optimal choice of ABC II approach is problem dependent. A larger simulation study is required to determine which approach is better under different circumstances. Unfortunately, such a study would be very time consuming due to the computationally intensive nature of pBII. We suggest trying all three methods (assuming that ABC IP and ABC IL are computationally feasible). One approach to speed up ABC IP and ABC IL might be to start with a computationally simple but consistent estimator (e.g. the method of moments) and apply one iteration of a Newton-Raphson method to produce an asymptotically efficient estimator (Cox and Hinkley, 1979, pp. 308) in a timely manner.

From a practical perspective, these methods have led to improved approximate analyses for two substantive problems compared with that obtained in Drovandi et al. (2011) and Drovandi and Pettitt (2011). Across applications considered in this paper, ABC IS was the most computationally efficient and led to good posterior approximations.

Overall, pdBIL avoids having to choose an ABC discrepancy function and the ABC tolerance. If an auxiliary model can be proposed that satisfies a rather strong condition, more precise inferences can be obtained by taking n large, which we showed is still computationally feasible with MCMC pdBIL up to a point. However, ABC II appears to provide a more general framework for pBII problems, due to the extra flexibility of being able to incorporate additional summary statistics outside the set formed from the auxiliary model and potentially providing better approximations when the auxiliary model is a simplified version of the generative model. It is this extra flexibility that may see ABC II as the method of choice as ever-increasingly complex applications are encountered.

Acknowledgements

The first two authors were supported by the Australian Research Council. The authors would like to thank Edwin Michael and David Denham for access to the macroparasite data. The first author is grateful to Ewan Cameron, Richard Everitt, Dennis Prangle, Andy Wood and Christian Robert for useful discussions about this work. The first author is also grateful to Warwick University (where this work was partially completed) for providing some funding for a three month visit.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons, 2nd edition.

Blum, M. G. B. (2010). Approximate Bayesian computation: a non-parametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208.

Cox, D. R. (1961). *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, volume 1, chapter Tests of separate families of hypotheses, pages 105–123. University of California Press, Berkeley.

Cox, D. R. and Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Cox, D. R. and Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika*, 77(4):747–761.

Creel, M. D. and Kristensen, D. (2013). Indirect likelihood inference.

Davison, A. C. (2003). *Statistical models*, volume 11. Cambridge University Press.

Denham, D. A., Ponnudurai, T., Nelson, G. S., Guy, F., and Rogers, R. (1972). Studies with *Brugia pahangi*. I. Parasitological observations on primary infections of cats (Felis catus). *International Journal for Parasitology*, 2(2):239–247.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B. Methodological*, 46(2):193–227.

Drovandi, C. C. (2012). *Bayesian algorithms with applications*. PhD thesis, Queensland University of Technology.

Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55(9):2541–2556.

Drovandi, C. C., Pettitt, A. N., and Faddy, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):503–524.

Drovandi, C. C., Pettitt, A. N., and McCutchan, R. A. (2013). Exact and approximate Bayesian inference for low count time series models with intractable likelihoods. *http://eprints.qut.edu.au/60032/*.

Gallant, A. R. and McCulloch, R. E. (2009). On the determination of general scientific models with application to asset pricing. *Journal of the American Statistical Association*, 104(485):117–131.

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory*, 12(04):657–681.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gleim, A. and Pigorsch, C. (2013). Approximate Bayesian computation with indirect summary statistics. Technical report, University of Bonn.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118.

Heggland, K. and Frigessi, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 66(2):447–462.

Jiang, W. and Turnbull, B. (2004). The indirect method: inference based on intermediate statistics - a synthesis and examples. *Statistical Science*, 19(2):239–263.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Michael, E., Grenfell, B. T., Isham, V. S., Denham, D. A., and Bundy, D. A. P. (1998). Modelling variability in lymphatic filariasis: macro filarial dynamics in the *Brugia pahangi* cat model. *Proceedings of the Royal Society of London: Series B*, 265:155–165.

Murray, I., Ghahramani, A., and MacKay, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia. AUAI Press.

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75.

Reeves, R. W. and Pettitt, A. N. (2005). A theoretical framework for approximate Bayesian computation. In Francis, A. R., Matawie, K. M., Oshlack, A., and Smyth, G. K., editors, *Proceedings of the 20th International Workshop on Statistical Modelling*, pages 393–396, Sydney, Australia.

Riley, S., Donnelly, C. A., and Ferguson, N. M. (2003). Robust parameter estimation techniques for stochastic within-host macroparasite models. *Journal of Thoretical Biology*, 225(4):419–430.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Smith, Jr., A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(S1):S63–S84.

Suswillo, R. R., Denham, D. A., and McGreevy, P. B. (1982). The number and distribution of *Brugia pahangi* in cats at different times after primary infection. *Acta Tropica*, 39(2):151–156.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.

Appendix A

We let

$$p(y|\theta) = \mathcal{N}(y; \theta, 1)$$

and so the ABC artificial likelihood, with $K_\epsilon(\rho(s(x_{1:n}), s(y))) = \mathcal{N}(\rho(s(x_{1:n}), s(y)); 0, \epsilon)$ and $s(x_{1:n}) = \frac{1}{n}\sum_{j=1}^{n} x_i$ is

$$p_{\epsilon,n}(y|\theta) = \mathcal{N}(y; \theta, n^{-1} + \epsilon)$$

and if the prior on θ is $p(\theta) = \mathcal{N}(\theta; 0, \sigma^2)$ then the posterior is

$$\begin{aligned} p_{\epsilon,n}(\theta|y) &= \mathcal{N}\left(\theta; \frac{y}{(n^{-1}+\epsilon)} \cdot \left[\frac{1}{n^{-1}+\epsilon} + \frac{1}{\sigma^2}\right]^{-1}, \left[\frac{1}{n^{-1}+\epsilon} + \frac{1}{\sigma^2}\right]^{-1}\right) \\ &= \mathcal{N}\left(\theta; \frac{y\sigma^2}{\sigma^2 + n^{-1} + \epsilon}, \frac{\sigma^2(n^{-1}+\epsilon)}{\sigma^2 + n^{-1} + \epsilon}\right) \end{aligned}$$

whereas

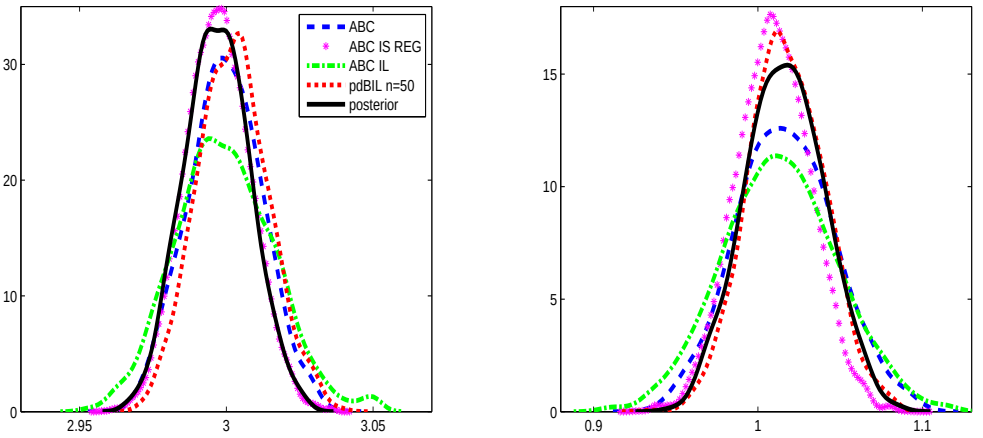
$$\begin{aligned} p(\theta|y) &= \mathcal{N}\left(\theta; y \cdot \left[1 + \frac{1}{\sigma^2}\right]^{-1}, \left[1 + \frac{1}{\sigma^2}\right]^{-1}\right) \\ &= \mathcal{N}\left(\theta; \frac{y\sigma^2}{\sigma^2+1}, \frac{\sigma^2}{\sigma^2+1}\right) \end{aligned}$$

In this example the posterior mean moves closer to y and has a variance tending to 0 as $n \to \infty$ and $\epsilon \to 0$.

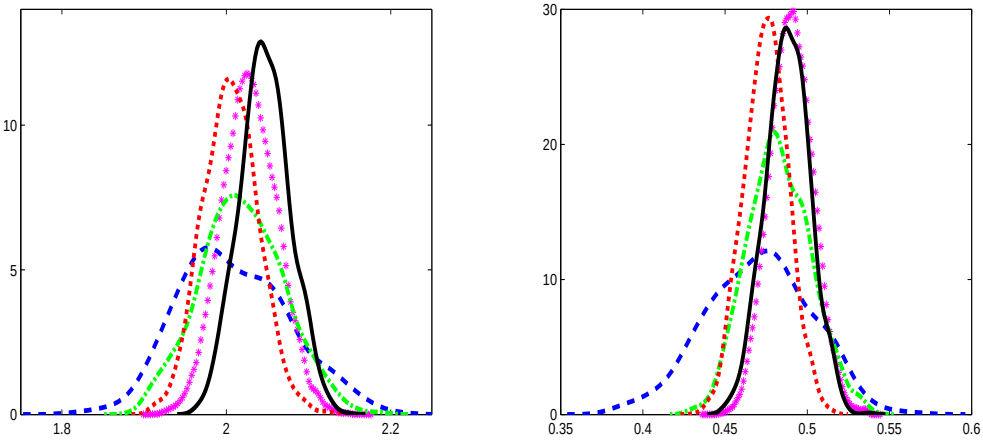
Appendix B

Applying the same process as in the main paper, a four component auxiliary mixture model produced a p-value of roughly 0.26, indicating a much better fitting auxiliary model. The BIC of the four component mixture model was 32114 compared to 32401 for the three component mixture model. The superior fit of the four component model would indicate that summary statistics drawn from the auxiliary model would carry more of the information of the observed dataset. However, an extra component creates three additional summary statistics, which could create issues for ABC II. Furthermore, the EM algorithm applied to four components was much more expensive and the algorithm was more prone to getting stuck in local modes. In terms of the BIC, it was found that a five component mixture model fits best, but using this was not computationally feasible.

The results comparing the pBII when the four component mixture auxiliary model was used is shown in Figure 11. Overall, the ABC IS with regression adjustment provided the best approximation. Further, the ABC IS approach is able to avoid the heavy computation associated with fitting the four component mixture model at every iteration and thus also avoids other numerical issues such as the EM algorithm converging to potentially local optima.



(a) a (b) b



(c) g (d) k

Figure 11: Posterior distributions for the parameters of the g-and-k model based on the ABC approach of Drovandi and Pettitt (2011) (blue dash), ABC IS with regression adjustment (purple star), ABC IL (green dot-dash) and pdBIL with $n = 50$ (red dot) and results from using the true likelihood (black solid). Here a four component normal mixture model was used as the auxiliary model (on-line figure in colour).

Appendix C

Here we describe the results for the pBII methods when applied to the simulated macroparasite data of the main article. To determine the goodness-of-fit of the auxiliary model (we use the same auxiliary model specified in the main article) to this data, we considered the following based on the generalised Pearson test statistic

$$T(\boldsymbol{\phi}) = \sum_{i=1}^{N} \frac{(m_i - \mu_i(\boldsymbol{\phi}))^2}{V_i(\boldsymbol{\phi})},$$

where $\mu_i(\boldsymbol{\phi})$ and $V_i(\boldsymbol{\phi})$ denote the mean and variance of the beta-binomial model for observation i, respectively, which depend on the auxiliary parameter $\boldsymbol{\phi}$. For the beta-binomial model, these are given by

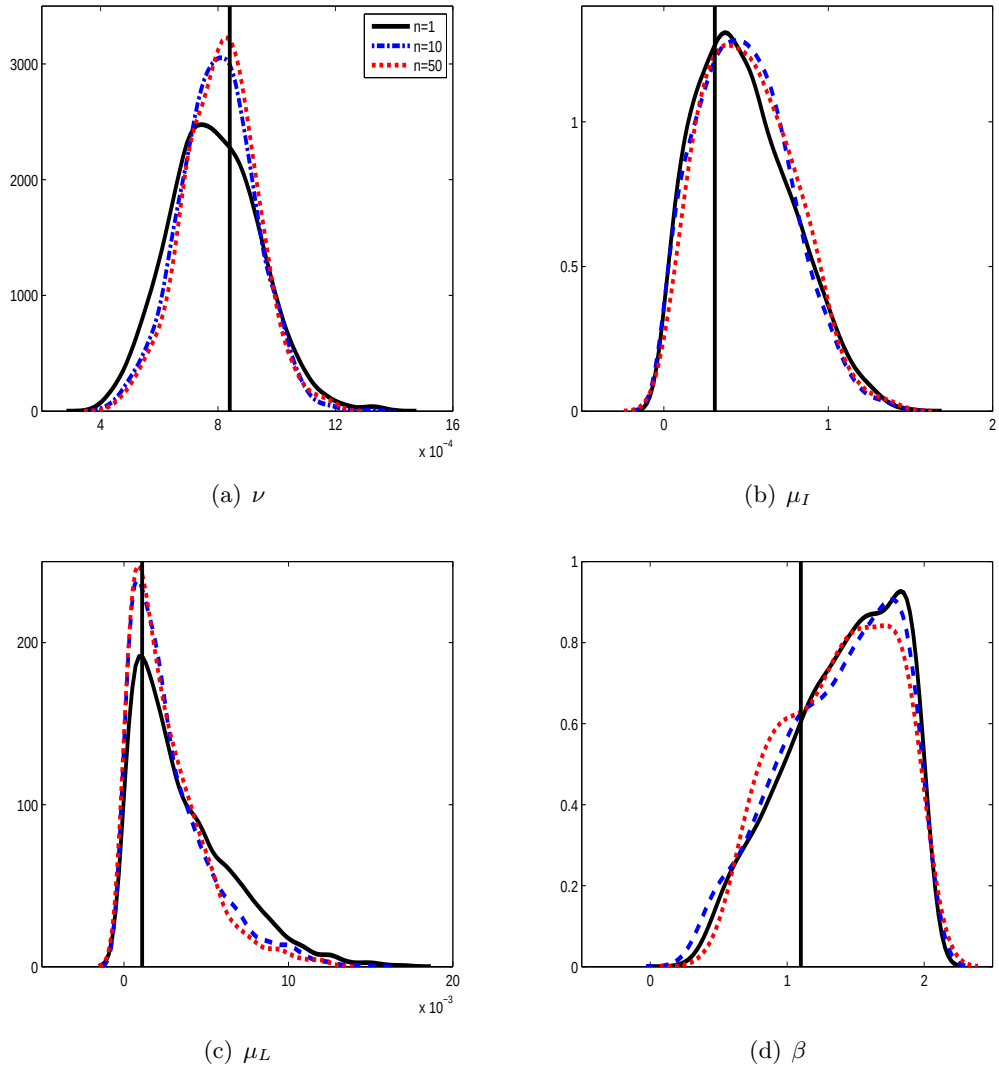
$$\mu_i = \frac{l_i \alpha_i}{\alpha_i + \beta_i},$$
$$V_i = \frac{l_i \alpha_i \beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)},$$

where the dependence of μ_i, V_i, α_i and β_i on $\boldsymbol{\phi}$ has been dropped for notational convenience. Here we found that $T(\boldsymbol{\phi}(\boldsymbol{y})) = 229.1$. Assuming that this test statistic has a chi-square distribution with degrees of freedom $N - 5$, the p-value obtained was roughly 0.14. To avoid making a distributional assumption about the test statistic, we performed a hypothesis test via Monte Carlo simulation where the null hypothesis was that the observed data can be described by a beta-binomial distribution with parameter $\boldsymbol{\phi}(\boldsymbol{y})$. Based on 1000 Monte Carlo draws from the null hypothesis, we found that the simulated test statistic values exceeded that of the observed test statistic about 18% of the time. Either way, there is insufficient evidence against the assumption that the beta-binomial distribution provides a good description of the data.

Figure 12 shows the results for pdBIL for $n = 1$ (100000 iterations), $n = 10$ (50000 iterations) and $n = 50$ (10000 iterations) using a suitable thinning interval. The acceptance probabilities were roughly 14%, 32% and 39% for $n = 1$, $n = 10$ and $n = 50$ respectively. In all cases, about 3 hours of computing time was required. Here we see an increase in precision for ν and μ_L as n is increased.

Figure 13 includes results for the ABC II methods. ABC IP and ABC IL were run for 1 million iterations and the tolerance was tuned to obtain an acceptance probability of approximately 1.5%-2% (roughly 20 hours of computing each). ABC IS was run for 20 million iterations due to the available analytic score for the auxiliary likelihood (15 hours of computing and an acceptance rate of roughly 0.2%). It is evident from the graph that generally the pBII methods are able to recover ν and μ_L quite precisely, and μ_I but less so. Given that the observed data appear to quite uninformative about μ_I and β (see Drovandi et al. (2011)), this simulation study provides a validation of the pBII approaches. Of the pBII methods, the ABC IS method appeared to give the most precise results for ν and μ_L.

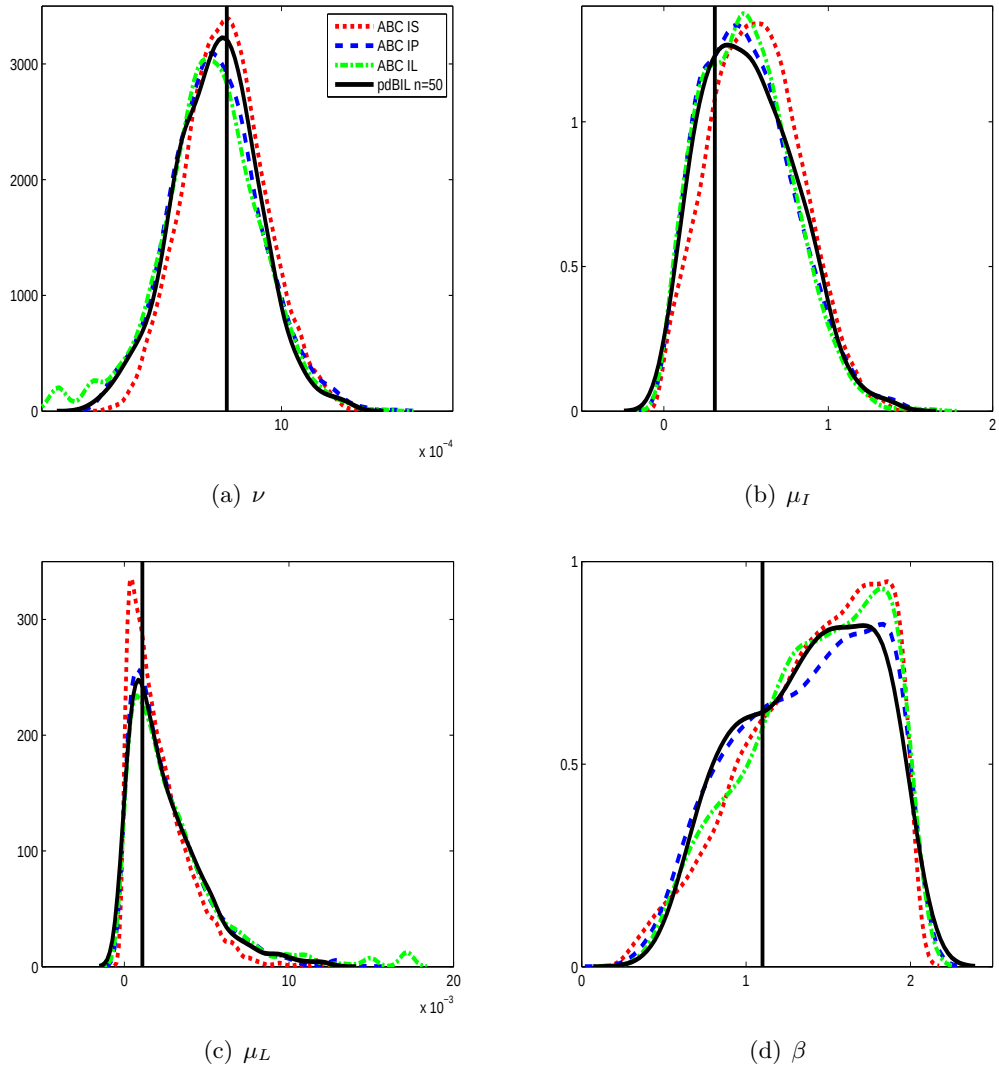
It is unclear if the performance of ABC IS is superior to the other ABC II methods due to the choices of the ABC tolerance. Here we applied a regression adjustment for the parameters $-\log(\nu)$ and $\sqrt{\mu_L}$ in an attempt to reduce the effect of the ABC tolerance. Figure 14 shows



(a) ν (b) μ_I

(c) μ_L (d) β

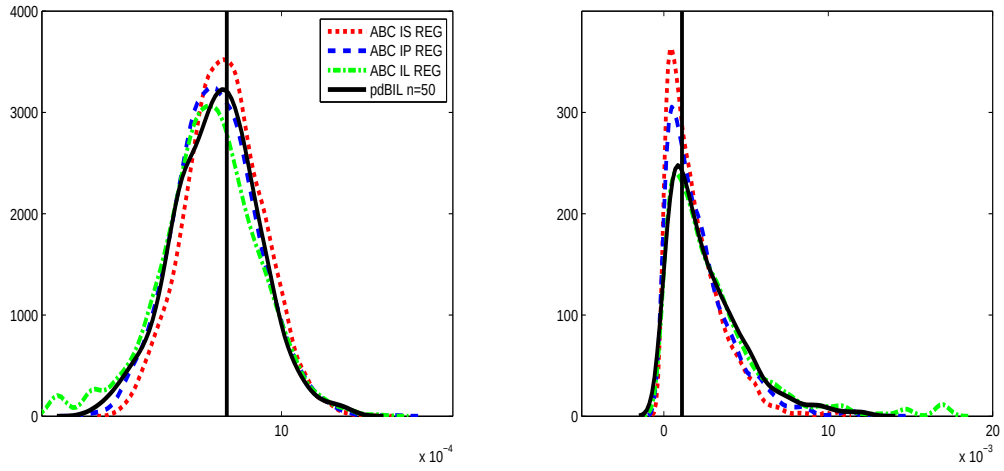
Figure 12: Posterior distributions for the parameters ((a) ν, (b) μ_I, (c) μ_L, (d) β) of the macroparasite model based on an pdBIL approach with $n = 1$ (black, solid), $n = 10$ (blue, dashed) and $n = 50$ (red, dotted) applied to the simulated data (on-line figure in colour).



(a) ν (b) μ_I

(c) μ_L (d) β

Figure 13: Posterior distributions for the parameters ((a) ν, (b) μ_I, (c) μ_L, (d) β) of the macroparasite model applied to the simulated data based on ABC IP (blue dash), ABC IS (red dot), ABC IL (green, dot-dash) and pdBIL (black solid) (on-line figure in colour).



(a) ν (b) μ_L

Figure 14: Posterior distributions for the parameters ((a) ν and (b) μ_L) of the macroparasite model based on ABC IP (blue dash), ABC IS (red dot), ABC IL (green dot-dash) and pdBIL (black solid) when applied to the simulated data. Regression adjustment has been performed on the output of the ABC II approaches (on-line figure in colour).

the results. We found slightly more precise inferences for the ABC IP and ABC IS results and negligible change for ABC IL. Overall, for this example based on this simulated dataset, it appears that ABC IS produced the best results. However, all of the pBII methods performed quite well.